Measuring Intra-Individual Change at Two or More Occasions With Hypothesis Testing Methods

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Chaitali Phadke

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. David J. Weiss, Adviser

December 20, 2017

Acknowledgements

I want to express my sincere gratitude and thanks to my adviser and mentor, Dr. David J. Weiss. This dissertation would not have been conceived or come to fruition without his expertise, motivation, guidance, and confidence in me. Thank you for your continuous support and patience. You have been, and will always be, a source of inspiration to me. Thank you Dr. Weiss for being an excellent academic adviser that every PhD student wishes for!

Many thanks to my committee members – Dr. Waller, Dr. Wang and Dr. Christ – for agreeing to serve on this dissertation committee. The input and feedback you provided were immensely helpful in improving my own understanding of the subject matter and the quality of my work. Many thanks to Dr. Christ for providing the K-12 data which was a great value addition to my dissertation. My special thanks to Dr. Waller, my very first contact at the University of Minnesota (UMN), who initiated me into my graduate life at the UMN. Many thanks to Dr. Wang as well whose input proved to be valuable in conceptualizing my research problem. Many thanks to Susan, my manager at Scantron for being a great supervisor and giving me the flexibility that allowed me to speed up my dissertation while doing justice to my office work.

Having a lovely cohort – Lian, Leah – and a wonderful set of colleagues at the Psychometrics Program at the UMN – Steve, Jeff, Jieun, Chris, Joy, Alec, Shengyu, Justin, Ziming, Lauren – has only made my graduate life more enjoyable. Thanks to Hongyuan for his help with programming expertise.

My graduate life would not have been easy without Pooja, Prerna, Shreyas, Sid, Asavari, Iravati, Sanju, Shekhar, Gnanada, Sameer, Pinal, Mehak all my friends who became the second family.

Pursuing a life and career in a country far away from home would not have been possible without my parents' – Ravikiran and Sugandha's – never ending encouragement and their belief in me. Thank you for continuous love, support and sacrifices you have made to shape up my life. These last months of thesis writing would not have been as easy without your physical, mental, and emotional support! Thank you for putting up with my tantrums and frenzies with happy faces. Also sincere thanks to my other set of family: my parents-in-law Ajitkumar and Arundhati who always encouraged me in my academic pursuits. Thanks to Asavari and Sandeep for pampering me during the break between the rigorous semesters!

I cannot thank Vratesh, my friend, confidant and now husband, enough. What did he not do for me? From far and near, he supported me in every possible way – he became my punching bag when I was frustrated and a cheerleader when I was down, he pushed me when I became lazy or procrastinated and had immense faith in me, at times more than I had in myself. Thank you for all the love, positivity, and happiness that you bring into my life!

I dedicate this thesis to my loving grandmas – Malti and Susheela – who were my friends, mentors, guides, philosophers and much more. I miss you a lot; how I wish you could see me achieve this academic milestone! But I take solace in a belief that you will continue to shower your blessings from wherever you are.

Abstract

The present study proposed six new omnibus hypothesis tests – F1, F2, LR, ST, χ^2_{FI} and χ^2_{GD} – to measure psychometric significance of individual change when an individual is measured at two or more occasions. The hypothesis tests were evaluated on criteria of Type I error, power, and agreement between the methods in the adaptive measurement of change (AMC) framework. This study expanded on AMC research by Finkleman, Weiss and Kim-Kang (2010) and Lee (2015), by introducing more generalized methods for the multi-occasion case. The omnibus tests were evaluated under various discrimination, bank type, and change conditions. The simulation results showed the LR test to achieve an optimum balance between Type I error and power. The hypothesis tests were found robust under most testing conditions. The tests were successfully applied to K-12 math data. The proposed methods are applicable under a variety of testing conditions in which IRT-based item parameters have been established.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In many fields of psychology, individuals are measured repeatedly over a period of time. For example, in a clinical setting a therapist who is interested in testing whether the clinical symptoms have reduced in a patient may measure the patient before and after therapy. In an academic setting, teachers may want to gauge students' understanding of the subject matter over a period of an academic year. In an industrial setting, human resources professionals might be interested in assessing the effectiveness of a training program as reflected in performance on the task by the employees. The point of interest in such measurement is an individual's growth over a period of time as a result of the intervention.

While there has been more emphasis on group level change in the statistics and psychometric literature (Arriaga, 1984; Dimitrov & Rumrill Jr., 2003; Guyatt, Walter & Norman, 1987; McArdle & Epstein, 1987), various methods of measuring individual change have also been proposed (Burr & Nesselroade, 1990; Finkleman, Weiss & Kim-Kang, 2010; Manning & Du-Bois, 1962; McDonald, 1999). Measurement of individual change has been a topic of controversy in the psychometric literature (Bereiter,1963; Cronbach & Furby, 1970; Embretson, 1995; Harris, 1963; Williams & Zimmerman, 1996; Zimmerman & Williams, 1982). Most of this criticism concerns the inability of classical test theory (CTT) based methods to measure change reliably using difference scores. Cronbach and Furby (1970) even suggested that researchers should abandon measuring change, or frame the questions related to change differently.

Presented below is a review of methods of measuring individual change. The methods can be broadly classified into CTT based methods and item response theory (IRT) based methods.

Classical Test Theory Based Methods

Difference Score

One of the most traditional approaches to measuring individual change is to compute a difference score between the measurements. The simple difference score is given by

$$D_j = Y_j - X_j, \tag{1}$$

where D_j is the observed change or difference score for person j, Y_j is the observed score at Occasion 2, and X_j is the observed score at Occasion 1. Previous research has demonstrated that use of simple difference scores is questionable for a number of reasons. First, simple difference scores tend to have lower reliability than the component variables (Bock, 1976; Embretson, 1995; Hummel-Rossi & Weinberg, 1975; Lord, 1963; Overall & Woodward, 1975; Willett, 1994, 1997). In CTT, reliability of a measure is defined to be a ratio of true score variance to total score variance (Crocker & Algina, 2006). Reliability of difference scores, is then

$$\rho_{DD'} = \frac{\sigma_{DD'}}{\sigma_D^2} = \frac{\sigma_X^2 \rho_{XX'} + \sigma_Y^2 \rho_{YY} - 2\sigma_X \sigma_Y \rho_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X \sigma_Y \rho_{XY}}, \tag{2}$$

where $\rho_{DD'}$ is the reliability coefficient of the difference scores, $\rho_{XX'}$ is the reliability of scores at Occasion 1, $\rho_{YY'}$ is the reliability of scores at Occasion 2, σ_X^2 is the variance of scores at Occasion 1 and σ_Y^2 is the variance of scores at Occasion 2. Assuming equal variance and reliability for *X* and *Y*, the above equation can be simplified to

$$\rho_{DD'} = \frac{\rho_{XX'} - \rho_{XY}}{1 - \rho_{XY}}. \quad (3)$$

When the scores at Occasion 1 and Occasion 2 covary in the same direction (which is often the case), then the difference scores tend to be less reliable ($\rho_{DD'}$) than the measure itself ($\rho_{XX'}$). Furthermore, change scores tend to have a negative correlation with an individual's initial status/score (Cronbach & Furby, 1970; Embretson, 1995; Rogosa & Willett, 1983; Willett, 1994, 1997). Because of ceiling effects, individuals with low scores at Occasion 1 are likely to show more change compared to individuals with high scores at Occasion 1 (Zimmerman & Williams, 1982). This implies that any variable that is positively correlated with Occasion 1 scores will have an artificial negative correlation with the difference score (Markus, 1980). Lord (1963) showed that the negative correlation of the difference score with the initial score came from the tendency of regression toward the mean from pretest to posttest measurement. This regression effect may be attributed to the fact that the individuals in the sample change at different rates: examinees who obtain low scores at Occasion 1 may show more improvement at Occasion 2 than those who obtain high scores at Occasion 1 (Bohrnstedt, 1969). Use of raw difference scores has also been criticized on the grounds that raw scores do not adequately represent the actual ability that underlies the performance on a (pre- or post-) test. In general, the relationship between raw scores and gain scores is not linear. Hence, equal gain scores may not represent equal change in ability (Dimitrov & Rumrill, 2003). Fischer (1976) demonstrated that, if a low ability person and a high ability person have made the same change on a particular ability scale (i.e., derived exactly the same benefits from the treatment), the raw-score differences will be misleading. Specifically, with a relatively easy test, the raw-score differences will (falsely) indicate

higher change for the low ability person and, conversely, with a more difficult test, they will (falsely) indicate higher change for the high ability person.

Residual Change Score

The residual change score (RCS), proposed by Manning and Dubois (1962) is one of the most commonly used alternatives to the simple difference score (Willett, 1997). Manning and Dubois (1962) demonstrated that the RCS is more reliable than the difference score. The RCS is given by

$$R_j = Y_j - Y_j' \tag{4}$$

$$R_j = Y_j - \overline{Y} - b_{Y.X}\left(X_j - \overline{X}\right), \tag{5}$$

where Y_j is the observed score at Occasion 2 for person j, X_j is their observed score at Occasion 1, Y_j' is the predicted score from the bivariate regression of Y on X, $\overline{Y}$ and $\overline{X}$ are the means of distributions of observed scores at Occasion 2 and Occasion 1, respectively, and $b_{Y.X}$ is the slope of the linear regression of Y on X. The residual change score is not a direct measure of gain/loss, but reflects a difference between the observed and the predicted score at Occasion 2 based on the Occasion 1 score for person j. Group level information is necessary for obtaining the RCS. R_j is the measure of difference between observed and predicted change score on the basis of a simple linear regression model. R_j in itself does not provide information about the magnitude or the implications of the observed change. Hence, the measure is appropriate for studying the correlates of change, but not for evaluation of individual change (Kim-Kang & Weiss, 2008).

Reliable Change Index

The reliable change index (*RC*) was introduced by Jacobson, Follette and Revenstorf (1984). It is defined as

$$RC = \frac{D}{SEM}, \tag{6}$$

where *D* is the difference score between pretest and posttest and *SEM* is the standard error of measurement, i.e., $SEM = \sigma_X\sqrt{1-\rho_{XX'}}$. Different authors have used different estimates of measurement error in the denominator of the reliable change index. Christensen and Mendoza (1986) proposed that the standard error of the difference between two test scores (SEM_D) should be the appropriate error term. SEM_D equals the standard deviation (SD) of the individual's hypothetical change-score distribution. Assuming equal *SEM* for pretest and posttest, the result is $SEM_D = \sqrt{2}SEM$ (e.g., Maassen, 2004). Assuming that the *RC* index follows a standard normal distribution, it can be concluded that there is a significant change if | *RC* | is more extreme than the $1-\alpha/2^{th}$ quantile of the standard normal distribution (Kruyen, Emons & Sijtsma, 2014). Kruyen et al. (2014) studied performance of the *RC* index under different conditions including varying number of items and amount of true change. Detection rates were measured as proportion of simulees identified as showing change by the *RC* index. Kruyen et al. (2014) reported the detection rate of the *RC* index to be around 0.9 for long tests ($n = 40$) and change of 1.5 standard deviations. Detection rate ranged from 0.4 to 0.7 for short tests ($n = 10$, 15 or 20). Detection rates were reported to be very low, around 0.15 to 0.40 for change of 0.5 and 1.0 standard deviations.

Minimal Important Difference

The minimal important difference (MID) is the smallest true score change between pretest and posttest that is perceived to be important by a clinician who is an expert in a client's functioning (Bauer, Lambert, & Nielsen, 2004; Kruyen, Emons & Sijtsma, 2014; Norman, Sloan, & Wyrwich, 2003; Schmitt & Di Fabio, 2004). If an individual's change

score is smaller than MID, the change can be considered practically unimportant. MID is often defined as a half standard deviation of the total score distribution (Norman et al., 2003). However, some authors (Jacobson et al., 1984; Wise, 2004) have argued that this approach is inadequate due to the unreliability of change scores and the presence of measurement error. Kruyen, Emons and Sijtsma (2014) have recommended that individual change be considered to be clinically important if | D | > MID and | D | differs significantly from 0. | D | is said to differ significantly from 0 if | RC | is more extreme than the $1 - \alpha/2^{\text{th}}$ quantile of the standard normal distribution. A clinician can even be more conservative by deciding that a client showed a clinically important change only if | D | is significantly larger than MID. This implies testing H_0: | T_D | ≤ MID against alternative H_1: | T_D | > MID, where T_D denotes the change parameter under the null hypothesis.

Item Response Theory Based Methods

Most methods of measuring individual change have been restricted by the use of instruments which are constructed on the basis of CTT. The quality of the measurement suffers when such tests are used for repeated testing. This problem, however, has been rarely discussed in the psychometric literature (Embretson, 1996; Kang & Waller, 2005; Von Minden, 2011; Weiss & Von Minden, 2011). When the response is in binary format, tests constructed using CTT typically contain highly discriminating items with difficulty around 0.5 so as to maximize the test reliability (Crocker & Algina, 2006), a scenario more common in the educational than the personality domain of measurement. Such tests tend to be peaked with a narrow range of item difficulty. When individuals are measured repeatedly with either the same test or some parallel form of it, the properties of the measure constructed on the basis of CTT remain the same. However, if there is a change

in the latent trait as a result of time or an intervention, the same CTT-based instrument now becomes less useful for detecting change. The instrument becomes "off-target" as a result of the shift in the latent trait. Hence, there is much larger measurement error at the later times of the testing. Measuring individual change with CTT is further hampered by the fact that item statistics in CTT, i.e., item discrimination and item difficulty, are sample based.

Item response theory (IRT) provides several advantages over CTT in measuring individual change. In terms of the development of measuring instruments, IRT has been replacing CTT (Embretson & Reise, 2000). With IRT, examinees can be placed on the same scale not only when different examinees are measured using different items, but also when the same examinees are measured repeatedly across time. When tests are constructed using IRT, items from a broad range of difficulty can be included to provide information over a range of ability. "This is plausible because item parameters estimated using one sample of examinees at a certain level of latent trait can be linked or transformed to those with another level of the trait, thereby allowing the development of item banks that can cover a wide trait range" (Finkleman, Weiss, & Kim-Kang, 2010). Thus, with IRT, tests can be built along different levels of the trait continuum. IRT, therefore, provides an advantage over CTT in measuring individual change. Jabrayilov, Emons, and Sijtsma (2016) compared CTT and IRT based methods to detect change with respect to Type I error and power. They found that IRT based methods were superior to CTT based methods in individual change detection, provided that the tests consist of at least 20 items. For shorter tests, however, CTT was generally better at correctly detecting change in individuals. The following section describes IRT based approaches of measuring change. The IRT models for measuring change are also classified as "Longitidinal IRT Models" by several authors

(e.g., De Boeck& Wilson, 2004; Fox & Glas; 2001; McArdle et al., 2009; Wang, Kohli & Henn, 2015).

Linear Logistic Model

Fischer (1976, 1983) developed the linear logistic model for measuring change within the framework of the generalized Rasch model. For Occasion 1, the probability that person *j* responds to item *i* correctly is defined as

$$P(\theta_{ij}) = \frac{\exp(\theta_{ij} - b_i)}{1 + \exp(\theta_{ij} - b_i)}, \tag{7}$$

where θ_{ij} is ability for person *j* associated with item *i* and b_i is the item difficulty associated with item *i*. At Occasion 2, *m* treatments that are applied to person *j* are accounted for in the model, so that the probability of correctly responding to item *i* is defined as

$$P(\theta_{ij}, \eta_m) = \frac{\exp(\theta_{ij} - b_i + \sum_{k=1}^{m} q_{jk}\eta_k + \tau)}{1 + \exp(\theta_{ij} - b_i + \sum_{k=1}^{m} q_{jk}\eta_k + \tau)}, \tag{8}$$

where $q_{jk}, \ldots, q_{jm}$ are doses of *m* treatments, η_k is the effect for treatment *k*, and τ is the trend effect, which is independent of the treatments (e.g., natural maturation). Fischer's logistic linear model is appropriate for group level of change, but not for measuring individual change, because the treatment effects and trend effects are assumed to be constant for everyone across all occasions (Lee, 2015).

Anderson's Rasch Model for Repeated Administration

Anderson (1985) proposed a Rasch model for the repeated administration of the same items over occasions. Anderson's model has the following form

$$P(\theta_j) = \frac{\exp(\theta_{jk} - b_i)}{1 + \exp(\theta_{jk} - b_i)}, \tag{9}$$

where θ_{jk} is the ability of person j at occasion k and b_i is the difficulty for item i. In Anderson's model, item difficulties are constant over occasions, but the ability that is involved depends on the occasion. Thus, occasions are characterized by different abilities, which may be correlated. Although Anderson's model is appropriate for understanding the impact of time or treatment on the ability distribution, the model does not contain change parameters for individuals. Abilities in Anderson's model are occasion-specific, and do not reflect person differences in changes over occasions (Embretson, 1991). Andrade and Tavares (2005) extended Anderson's model into a three-parameter logistic (3PL) parameterization. They allowed θ_{jk} to follow a multivariate normal distribution, so that serial correlations among θ_{jk}s were captured by a covariance matrix.

Multidimensional Rasch Model for Learning and Change

Embretson (1991) has proposed a multidimensional Rasch model for learning and change (MRMLC) for repeated measurements. In the MRMLC, performance on the k^{th} occasion is assumed to be a function of k abilities (i.e., ability at each of k occasions). The probability that person j responds to item i correctly under k occasions is defined as

$$P(\boldsymbol{\theta}_j) = \frac{\exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}{1 + \exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}, \tag{10}$$

where $\boldsymbol{\theta}_j$ is the vector of abilities in which θ_{jm} is the ability at $k = m$, and b_i is the item difficulty. The response for item i administered at occasion k is a function of all θs up to k occasions. The MRMLC measures individual change; however, it is restricted to the unrealistic assumption of equal item discriminations. The model complexity also increases as more measurements are taken: when repeated measurement is used, more person parameters are estimated at later occasions with the same set of items.

The MRMLC model was further extended to a two-parameter logistic (2PL) version, namely, the structured latent trait model (SLTM), by Embretson (1997). Although the MRMLC model does not require the same items to be used repeatedly to link the scales at different occasions, some of the items still need to be repeated over time (von Davier, Xu, & Carstensen, 2011).

Item Response Change Model

Mellenbergh and van den Brink (1998) proposed an item response change model also based on the Wiener simplex. In their item response change model, the probability of person *j* giving a correct answer to the i^{th} item at the k^{th} occasion is defined as

$$P_{ijk} = \frac{\exp\left(\gamma'_{ij1} + \gamma'_{j2} + \cdots + \gamma'_{jk}\right)}{1 + \exp\left(\gamma'_{ij1} + \gamma'_{j2} + \cdots + \gamma'_{jk}\right)}, \tag{11}$$

where γ'_{ij1} is the j^{th} respondent's expected response to the i^{th} item at the initial occasion, $\gamma'_{j2}, \ldots, \gamma'_{jk}$ are item change parameters for occasions 2, … , *k*, respectively, and are constant for all items per occasion. The maximum likelihood estimator of the change parameter is

$$\hat{\gamma}'_{j2} = \ln\left(n_p / n_n\right), \tag{12}$$

and the estimated variance is

$$Var(\hat{\gamma}'_{j2}) = \frac{n_p + n_n}{n_p n_n}, \tag{13}$$

where n_p and n_n are frequencies of test items changed in the positive (i.e., incorrect in the pretest and correct in the posttest) and negative (i.e., correct in the pretest and incorrect in the posttest) directions. However, their item change model only has an item-specific parameter γ'_{ij1} and this item-specific parameter is defined as the log odds of the j^{th}

respondent. Thus, the item change model in Equation 11 proposed by Mellenbergh and van den Brink (1998) is not appropriate for measuring individual change (Lee, 2015).

ANOVA Designs

The pretest-posttest data and the effect of the treatment as assumed to be reflected into varying pretest and posttest means is often investigated using the analysis of variance (ANOVA) designs (Bryk & Weisberg, 1977; Jöreskog & Sörbom, 1976; Linn, 1981; Linn & Slinde, 1977; Rumrill & Bellini, 2009; Sörbom, 1976; Stevens, 1996).

ANOVA on gain scores, analysis of covariance (ANCOVA) on gain scores, ANOVA on residual scores, and repeated measures ANOVA have been used traditionally in comparing groups with pretest and posttest data (Bryk & Weisberg, 1977; Jöreskog & Sörbom, 1976; Linn & Slinde, 1977; Rumrill & Bellini, 2009; Sörbom, 1976; Stevens, 1996). In simple ANOVA gain scores designs, gain scores are used as the dependent variable in comparison of two or more groups. In such designs, a null hypothesis of zero mean gain scores can be tested in a population of examinees. If the gain score is unreliable, however, it is not appropriate to correlate the gain score with other variables in a population of examinees (Mellenbergh, 1999). The ANCOVA design uses pretest scores as a covariate in pre-post data. When the design is randomized, ANCOVA serves to reduce error variance, because the random assignment of subjects to groups guards against systematic bias. With nonrandomized designs, the main purpose of ANCOVA is to adjust the posttest means for differences among groups on the pretest, because such differences are likely to occur with intact groups (Dimitrov & Rumrill, 2003). ANOVA on residual scores has also been used in the analysis of pretest-posttest data. The residual scores are the differences between observed and predicted post-test scores. Though residual scores contain less error than gain

scores (Zimmerman & Williams, 1982) and they do not correlate with the observed pretest scores, the ANOVA on residual scores is less powerful than that on the gain scores (Maxwell, Delaney & Manheimer, 1985). Maxwell, Delaney and Manheimer (1985) demonstrated that ANOVA on residual scores can result in an inflated α when the residuals are obtained from the pooled within-group regression coefficients. When the regression coefficient for the total sample of all groups combined is used, ANOVA on residual scores may also result in a conservative test. Repeated measures ANOVA is used with pretest-posttest data as a mixed (split-plot) factorial design with one between-subjects factor (the grouping variable) and one within-subjects (pretest-posttest) factor. However, the results provided by repeated measures ANOVA for pretest-posttest data can be misleading. Specifically, since the treatment does not affect the pretest scores, the F statistic for the treatment effect, which is of primary interest, can be conservative (Huck & McLean, 1975; Jennings, 1988) and may not detect true change often.

Assumptions such as randomization, linear relationship between pretest and posttest scores for the two or multi-occasion case, and homogeneity of regression slopes underlie most ANOVA models. In some cases, modification to the models is possible. For example, if there is a non-linear relationship between pretest and posttest scores, ANCOVA can be extended to include a quadratic or cubic component (Cahen & Linn, 1971). However, all the ANOVA models are useful for determining group level change, as they are essentially designed for comparing pretest-posttest data. They do not contain any individual-level parameter for treatment effect. Therefore, the ANOVA models are not useful in determining individual change.

Model-Based Approaches

Numerous model-based approaches of capturing growth have been proposed (Bryk & Raudenbush, 1987; Collins, 2006; Li, Cohen, Bottge & Templin, 2016; Macready & Dayton, 1977; Rogosa, Brandt & Willett, 1982; Rogosa & Willett, 1985; Strenio, Weisberg & Bryk, 1983; Ware, 1985; Waternaux, Laird, & Ware, 1985) in the literature. Collins and Sayer (2001) present an overview of contemporary developments in the field, with a focus on time series, dynamical, and multilevel models. The model-based approaches estimate the growth parameters based on longitudinal data with multiple observations rather than traditional pretest-posttest data.

Bryk and Raudenbush (1987) have proposed applying hierarchical linear models (HLM), one of the early applications of HLM, in assessing change. Latent growth curve models have also been popular in modeling change (Collins, 2006). Growth modeling approaches attempt to explain the structure of individual variability as well as change at the group level. In modeling latent growth trajectories, advanced statistical techniques such as hierarchical linear modeling, structural equation modeling, or time series have been applied (e.g., Bryk & Raudenbush, 1987; McArdle & Epstein, 1987; Espin, Deno & McConnell, 2004; Von Eye, 1990).

Bryk and Raudenbush (1987) described a two-stage hierarchical linear model. In their modeling approach, the first stage model consists of estimating individual growth parameters, as a function of individual growth trajectory. This is a within-subject stage. Systematic growth over time is represented as a polynomial of degree K – 1. Thus, the within-subject model is

$$Y_{it} = \pi_{0i} + \pi_{1i}a_{it} + \pi_{2i}a_{it}^2 + \cdots + \pi_{k-1i}a_{it}^{k-1} + R_{it}, \quad (14)$$

where Y_{it} is observed status of individual i at occasion t, a_{it} is age of individual i at occasion t, π_{ki} are growth trajectory parameters for individual i, and R_{it} is the random error assumed normally distributed with a mean of 0 and covariance structure Σ_i, dimensioned $T_i \times T_i$, for T occasions.

In the second stage of the model, i.e., between-subjects stage, the variation in growth trajectory (π_{ki}) between subjects is modeled as a function of background factors such as treatment/therapy/instruction effect, experimental treatment, and motivational factors. Specifically, each of the k individual growth parameters can be modeled as

$$\pi_{ki} = \beta_{k0} + \beta_{k1}X_{k1i} + \beta_{k2}X_{k2i} + \cdots + \beta_{kP-1}X_{kP-1i} + U_{ki}\,, \quad (15)$$

where there are $p = 1, \ldots, P-1$ measured variables (X_{kP}), β_{kP} represents the effect of X_{kP} on the k^{th} growth parameter, and U_{ki} is random error. U_{ki} are normally distributed with mean zero and covariance as

$$\text{cov}(U_{hi}, U_{ki}) = \text{cov}(\pi_{hi}, \pi_{ki}) = \tau_{hk} \quad (16)$$

for $h, k = 0, 1, \ldots, K-1$. The parameters β_{kP} are fixed effects and the errors U_{ki} are random effects. When the errors (R_{it}) are assumed to be independent with common variance σ^2, subject i's growth rate, π_i, can be estimated by means of ordinary least squares, based only on the repeated measurements for that subject, given by

$$\hat{\pi}_i = \Sigma a_{it} y_{it} / \Sigma a_{it}^2 \quad (17)$$

and the sampling variance of $\hat{\pi}_i$ for fixed π_i is

$$v_i = \text{var}(\hat{\pi}_i | \pi_i) = \frac{\sigma^2}{\Sigma a_{it}^2}. \quad (18)$$

When variances are unknown, the variance components must be estimated from the data. When number and spacing of the time series observations vary across subjects, variance estimation requires iterative, numerical approaches (Bryk & Raudenbush, 1987) such as the EM algorithm (Dempster, Laird & Rubin, 1977). Under fairly general conditions, the EM algorithm produces maximum likelihood estimates for variance components. These estimates have the desirable properties of being asymptotically unbiased, consistent, efficient, and asymptotically normally distributed (Bryk & Raudenbush, 1987; Dempster, Laird & Rubin, 1977). When the EM estimates are substituted for the unknown variances and covariances, the resulting β estimates are also maximum likelihood estimates with known asymptotic distributions. The latter provides the basis for large sample statistical inference with HLM (Bryk & Raudenbush, 1987; Dempster, Rubin, & Tsutakawa, 1981).

Li et al. (2016) combined a latent transition analysis (LTA; Collins & Wugalter, 1992) approach with a cognitive diagnostic model to account for change in the latent binary variables measured by cognitive diagnostic models. LTA is typically used for detecting the probabilities that members of different latent groups in the data will remain in those groups or shift into other latent groups. They addressed the question whether skill mastery statuses changed across the four test administrations and whether the changes subsequent to each instruction were similar or different by estimating probabilities in the transition matrix. These probabilities were calculated using a Markov Chain Monte Carlo (MCMC) algorithm employing Gibbs sampling. Using MCMC, they were able to identify frequencies and proportions of examinees who mastered each of the cognitive skills at each occasion. However, LTA can be particularly useful when used for investigating growth when latent variables are categorical (e.g., Boscardin, Muthén, Francis, & Baker, 2008; Compton,

Fuchs, Fuchs, Elleman, & Gilbert, 2008; Trentacosta et al., 2011). Although technically sophisticated, this model focuses on group level change and hence is not very useful in making inferences about individual change over multiple occasions of measurement.

Willett (1988-89) proposed linear, quadratic and exponential growth models for determining individual change by estimating slopes and intercepts for each individual. Estimation of these individual regression terms is based on the number of occasions a person is measured, also referred as multiwave data. Willett (1988-89) estimated the individual regression coefficients based on five or six occasion multiwave data. However, estimation and interpretation of the individual regression coefficients is not very useful. First, using the regression models based on six or seven multiwave data points would be inappropriate because regression coefficients derived from such small a number of observations would lead to erroneous interpretations. Second, the standard errors of such estimated coefficients would be large and hence interpreting them in terms of the magnitude of change would be unreliable.

Deriving from Willett's (1989) approach and using HLM to measure growth trajectories and its correlates, Shin et al. (2004) demonstrated the use of HLM and curriculum-based measurement (CBM) for assessing academic growth and instructional factors for students with learning difficulties. They demonstrated the use of HLM on CBM math data for students with and without learning disabilities. They used visual inspection of individual growth curves and used a deviance test (Bryk & Raudenbush, 1992) to select quadratic growth models to apply to the data.

In parallel with the previously cited work on IRT models for repeated measures designs, McArdle et al. (2009) proposed to combine IRT models with a longitudinal growth

curve (LGC) model. These models are classified under the broad category of multilevel models (e.g., De Boeck & Wilson, 2004; Fox & Glas, 2001). Wilson, Zheng, and McGuire (2012) proposed a latent growth item response model (LG–IRM), which allows for both linear and curvilinear growth patterns of the latent traits. Their model can also be viewed as a multidimensional IRT model (Wang & Nydick, 2015) constructed within the multidimensional random coefficient multinomial logit (MRCML; Adams, Wilson, & Wang, 1997) framework. LGC models were also extended into second-order LGC models which allow for modeling change in ability over time, where the latent factor is measured by multiple observed variables collected at each measurement occasion (e.g., Duncan, Duncan & Strycker, 2006; Hancock et al., 2001; Kohli & Harring, 2013; McArdle, 1988). Wang, Kohli and Henn (2015) demonstrated application of second-order LGC models for binary outcome variables in the IRT as well as structural equation modeling framework with detailed transformation equations to allow for different parameterizations.

The HLM approach allows for understanding systematic differences in individual growth trajectories as well as predict future development. However, like that proposed by Willett (1988-89), the estimations of individual change parameters are based on regression/ least-squares estimation from repeated measurements of a single subject. Such an approach may be useful when there are many measurements. However, in the case of three or four measurements, the estimates may be highly inaccurate. Also, if the distributional assumptions and assumptions about the covariances are not met, the HLM estimates may not be reliable (Bryk & Raudenbush, 1987). "Inferences based directly on the estimated variances and covariances need to be interpreted cautiously as these estimates depend heavily on the normality assumption and are also likely to be imprecise when sample sizes

are small. This means that the estimated correlation between initial status and rate of change and the estimated reliability of the growth parameters should be regarded as tentative when normality is questionable or if samples are small" (Bryk & Raudenbush, 1987, p. 156). Obviously, when these methods are applied to individual change the number of observations is the number of measurements taken, which are likely to be very small numbers. Another problem with using model-based approaches is that there is little or no consideration of whether the observed measurements actually reflect true change or whether significant change can be identified for a given individual. Thus, a different approach is needed that is designed to accurately measure and identify significant individual change when it exists.

Measuring Change with CAT

Computerized adaptive testing (CAT) is a type of computer-based assessment in which items are successively administered based on the performance of an examinee. Application of CAT has increased in various measurement domains (Fliege et al., 2005; Simms & Clark, 2005). CAT is composed of the following five components for a test administration (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011):

Components of CAT

Item Bank: A pre-calibrated item bank is required for CAT. This item bank is calibrated using IRT. Using IRT for CAT is particularly useful as examinees and items are on the same scale (Birnbaum, 1968). An item bank can be created to approximate a desired test information function (TIF). For a given IRT model, the item information function (IIF) for item *i* can be defined as

$$I_i(\theta) = \frac{[P_i(\theta)']^2}{P_i(\theta)Q_i(\theta)}, \tag{19}$$

where $P_i(\theta)'$ is the first derivative of $P_i(\theta)$ with respect to θ. $P_i(\theta)$ is the probability of a keyed response for item *i* given θ, which for the three-parameter logistic (3PL) IRT model is defined as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp[-Da_i(\theta - b_i)]}, \tag{20}$$

where $D = 1.7$, a scaling parameter. $P_i(\theta)$ in IRT is called an item response function (IRF). The TIF is the sum of item information available in a bank,

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{21}$$

where *n* is the number of items in an item bank. A CAT item bank is developed to obtain a TIF similar to a target TIF that is appropriate for the purpose of the test. For example, if the goal is to measure the examinees well across the θ level, then an item bank with a flat TIF can be created. If the goal is to measure examinees well around a certain θ level or a particular cut score, then a bank with peaked TIF measuring well around that θ level can be created.

Starting Point: In order to select the first item to be administered, one commonly used starting point for θ is a fixed value such as $\theta = 0$. However, this approach may lead to overexposure of items with difficulty at 0. In order to avoid such overexposure, items can be selected from a θ range. For example, -1 to 1. If information about the examinee's ability is available, such as a prior score from previous testing, or a score from a previous year, then such information can serve as a starting point as well.

Item Selection: In CAT, items are selected following an examinee response and a subsequent estimation of θ. A classical way to select an item is choosing the item from a non-administered item bank, providing maximum Fisher information in Equation 19 at the current θ estimate (Weiss, 1982). However, there are a couple of disadvantages of this approach. First, there is a danger of inaccuracy in θ estimates early in the test. Errors in the first few θ estimates are generally large in the early stages of CAT. Since the maximum information criterion selects items based on the current θ estimate, the performance of the selected item may not be optimal at the true ability level. Second, there is a risk of overexposure for highly discriminating items. Selecting an item with maximum discrimination at the current θ estimate may lead to concerns of test security and/or low bank utilization. The problems with maximum information item selection, including inaccurate θ estimate and overexposure of highly discriminating items, can be overcome by using one of several other item selection criteria (van der Linden & Pashley, 2010).

Scoring: θ is estimated in CAT after administration of each item. A maximum likelihood estimator (MLE) is commonly used as the θ estimation method. A likelihood function, $L(\theta|\mathbf{u})$, is obtained by multiplying IRFs of correct and incorrect responses for each item given θ,

$$L(\theta|\mathbf{u}) = \prod_{i=1}^{k} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}, \tag{22}$$

where k is the number of items administered so far in the test, $\mathbf{u}$ is a response vector, u_i is the i^{th} element of $\mathbf{u}$, and $Q_i(\theta) = 1 - P_i(\theta)$. The MLE is defined as the θ value that maximizes the likelihood function in Equation 22. The MLE has been reported to have a smaller bias function but larger standard error than those of Bayesian estimation methods

– maximum a posteriori (MAP) or expected a posteriori (EAP) (e.g., Wang, Hanson & Lau, 1999; Wang & Vispoel, 1998; Weiss, 1982). When MLE fails to converge (e.g., non-mixed item responses), alternative θ estimation methods such as MAP, EAP or weighted likelihood estimate (WLE) can be implemented.

Termination: CATs are terminated based on a pre-specified termination criterion. A fixed-length CAT, in which a fixed number of items has been decided upon, is terminated when the predetermined number of items has been administered. This is similar to a paper-pencil test. A variable-length CAT, in which such item limit is not specified, can be terminated using different criteria. For example, a CAT can be terminated when a certain level of precision is obtained in the θ estimate (Weiss & Kingsbury, 1984), or a minimal level of change is observed in the θ estimate (Gialluca & Weiss, 1979; Maurelli & Weiss,1981). An alternative method is to terminate the CAT when no items are left in the item bank that provide more than a minimal level of information at the current θ estimate (Hart et al., 2006; Weiss & Kingsbury, 1984). Variable-length CATs often impose constraints on a minimum and/or maximum test length to have practical justification for complaints from low performers with a test that is too short or to prevent all items in a bank from being administered (Thompson & Weiss, 2011).

Adaptive Measurement of Change

The adaptive measurement of change (AMC) approach was first used by Kingsbury and Weiss (1983). Kim-Kang and Weiss (2007, 2008) referred to CAT applied in measuring change as the adaptive measurement of change (AMC). The AMC approach uses CAT and IRT to obtain estimates of an individual's ability, ($\hat{\theta}$), from a domain of items. Similar to regular CAT, CAT banks are constructed at each measurement occasion

(and linked onto a common scale) according to the purpose of the tests, or the same bank with a high and wide range of item information is used. θs are estimated after each item is administered, and items are often selected to maximize the Fisher information at the current θ estimate. The tests are based on pre-specified termination criteria. For fixed-length termination, AMC is terminated after a pre-determined number of items are administered. AMC used at later occasions can be terminated when significant change is observed (i.e., variable-length termination). The estimates are obtained at different occasions and are separated by a certain amount of time. The difference in CAT θ estimates between the two or more occasions is defined as a measure of change. Measurement of change for a particular examinee is determined with reference to the previous trait level estimate.

CAT offers several advantages over conventional tests in measuring change. The problem of "off-target" testing (discussed above on pages 6 and 7 under the "Item Response Theory Methods" section) can be overcome using CAT over repeated measurements. With conventional tests, measurement precision decreases as an examinee's latent traits change. However, this may not be the case for CAT. Since CAT administers items that adapt to each examinee's θ level, CAT can provide precise measurement equally across θ levels at each occasion, and thus can provide a better estimate of change.

The problem of scale distortion (ceiling effect) in difference scores was investigated by May and Nicewander (1998) in the context of conventional and adaptive tests. They compared difference scores obtained from three different methods – difference in number-correct scores obtained by conventional tests, difference in θ estimates obtained by IRT scoring, and difference in θ estimates obtained by using adaptive tests. The results

indicated that the θ metric had smaller scale distortion compared to the number-correct score. Furthermore, even smaller scale distortion was observed for adaptive tests compared to the conventional IRT θ estimates. The difference scores used by May and Nicewander (1998) were in different metrics and they multiplied the number-correct difference scores by 10 for comparability with other IRT based and CAT methods. Moreover, the investigation focused on group level instead of individual level change. Hence, though not generalizable, the basic findings of their work showed superiority of CAT in measuring change.

Weiss and Kingsbury (1984) compared different methods of measuring individual change; namely simple difference score, the RCS, difference score based on IRT, and AMC. They found that the AMC method generally performed better than the other three methods in terms of capturing true individual change.

Kim-Kang and Weiss (2008) investigated performance of conventional tests and AMC in capturing true change. They compared number-correct score converted to the θ metric, RCS, IRT θ estimates and AMC for the two-occasion case. Weiss and Von Minden (2011) further extended the work of Kim-Kang and Weiss (2008) to five occasions with five growth curves including linear and curvilinear growth patterns. The recovery of individual true θ was evaluated based on bias and root mean square error of θ estimates at each occasion for each growth curve. The results of both studies showed that AMC consistently better estimated true growth with small error, while errors in conventional tests increased when tests were off-target of the examinee ability range. As was expected, the precision of measuring individual change decreased in conventional tests as examinees changed at later occasions.

Lee (2015) also reported that the AMC procedure demonstrated advantages over conventional tests in detecting true change. Using CAT, tests are designed to match an examinee's ability based on his or her previous answers. New items that provide more information at the examinee's ability level can be selected from an item bank, which makes the measurement more precise as well as more efficient (Finkleman, Weiss & Kim-Kang, 2010).

Hypothesis Testing in the Context of AMC

While the focus is on estimating the amount of change, most often practitioners also want to determine if the observed change is significant or not. For example, in a clinical setting, therapists are interested in knowing whether therapy has been effective in terms of bringing relief to the patients as reflected in reduced symptoms of depression (e.g., Falloon et al., 1985; Gagne & Toye, 1994; Smits et al., 2008). In an educational or academic setting, it is of great benefit to teachers to know whether students are learning or not and whether students who may be needing the most guidance are improving significantly, as reflected in their observed ability or achievement (e.g., Fennema et al., 1996; Lei & Zhao, 2007).

In the context of AMC, it can be determined if psychometrically significant change occurs within an examinee using hypothesis testing methods. We use the term "psychometric significance" instead of "statistical significance," as the hypothesis tests described in the context of AMC include error terms derived in the psychometric framework instead of the statistical sampling framework. The AMC hypothesis tests for measuring individual change do not use any group related information in estimation of the

test statistic, unlike in statistical tests. Hence, the significance of individual change is psychometrically more meaningful rather than being of statistical consequence.

The null hypothesis of no-change between two occasions of testing (H_0: $\theta_1 = \theta_2$) can be tested against an alternative hypothesis of H_0: $\theta_1 \neq \theta_2$. Individual change can be determined as psychometrically significant when the obtained test statistic is at least as extreme or more extreme than the $1 - \alpha/2^{th}$ quantile of the test statistic distribution. There have been a number of different methods proposed for determining significant change.

Confidence Intervals

Weiss and Kingsbury (1984) proposed constructing IRT-based confidence intervals around $\hat{\theta}$s. Significant change occurs when the confidence intervals estimated at each of the t = 2 testing occasions do not overlap. The confidence interval at occasion t is approximated as

$$CI = \hat{\theta}_t \pm z_{1-\alpha/2} \times SE(\hat{\theta}_t), \tag{23}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2^{th}$ quantile of the standard normal distribution and $SE(\hat{\theta}_t)$ is determined from the second derivative of the log likelihood function (Baker,1992, pp. 69–72; Weiss, 2005, pp. 10–11) with respect to θ,

$$SE(\hat{\theta}_t) = \frac{1}{\sqrt{-\delta^2 \ln L(\theta)/\delta\theta^2}}\Bigg|_{\theta=\hat{\theta}_t}, \tag{24}$$

where $L(\theta)$ is the likelihood function. $\hat{\theta}$ is a MLE which is determined from the maximum of the likelihood function (Equation 22). $\hat{\theta}$ is the value of θ at which the likelihood function is at its peak. This estimate is determined by an iterative Newton-Raphson procedure. The maximum of the function is that point at which the first derivative of the

likelihood function with respect to θ is 0, and the second derivative reflects the curvature of the function at its maximum. The iterative procedure continues until the ratio of the first derivative to the second derivative is arbitrarily small (De Ayala, 2009).

The confidence interval approach has been shown to be too conservative in detecting change, with lower than desirable Type I error as well as power. Kim-Kang and Weiss (2008) and Finkleman et al. (2010) demonstrated that when the desired Type I error level was set to 0.05, the observed Type I error based on this approach did not reach 0.01.

Using IRT, Fischer (2001, 2003) has applied Clopper-Pearson confidence intervals in testing the hypothesis of no-change $H_0: \lambda = 1$ against $H_a: \lambda \neq 1$. When change in Occasion 2 is expressed as $\theta_2 = \theta_1 + \eta$, where θ_1 and θ_2 indicate θ at Occasion 1 and Occasion 2, respectively, the parameter λ is defined as $\lambda = \exp(\eta)$. Thus, it becomes equivalent to testing $H_0: \eta = 0$ against $H_a: \eta \neq 0$. In Fischer's method, $\hat{\lambda}$ is a conditional maximum likelihood estimator, and confidence intervals are constructed using the conditional probability of posttest raw score given the sum of pretest and posttest total raw scores. This approach has been applied to conventional tests and is restricted to a family of Rasch models (i.e, dichotomous Rasch, Rasch rating scale, and Rasch partial credit models). Also, conditional maximum likelihood estimation and conditional probabilities in Fischer's application require item responses and item parameters from both pretest and posttest, which makes it difficult to apply in live adaptive testing.

Z test by Finkleman, Weiss, and Kim-Kang (2010)

A *Z* test for measuring individual change was described by Finkleman, Weiss, and Kim-Kang (2010). The *Z* statistic is presented as a standardized difference between θ estimates from two occasions.

$$Z_{FI} = \frac{|\hat{\theta}_2 - \hat{\theta}_1|}{\sqrt{I_2(\hat{\theta}_p)^{-1} + I_1(\hat{\theta}_p)^{-1}}}, \quad (25)$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates of θ at Occasion 1 and Occasion 2, respectively, and $I_t(\theta)$ denotes observed test information at Occasion *t*. The variance of $\hat{\theta}_2 - \hat{\theta}_1$ is estimated using test information from both occasions evaluated at $\hat{\theta}_p$, where $\hat{\theta}_p$ is the MLE of θ using the combined responses and item parameters from Occasion 1 and Occasion 2. The *Z* statistic is compared to a standard normal distribution to make a decision of significance. If the obtained statistic is at least as or more extreme than the $1 - \alpha/2^{\text{th}}$ quantile of the standard normal distribution, the change is determined to be psychometrically significant. This *Z* test has been demonstrated to have a Type I error around 0.05 (Finkleman et al. 2010; Lee, 2015) and power around 0.9 in conditions of medium (change of 1.0 SD) to high change (change of 1.5 SD).

***Z* test by Guo and Drasgow (2010)**

Guo and Drasgow (2010) also proposed a *Z* test in the context of detection of cheating in unproctored internet tests (UIT). They investigated change as the difference in the ability estimates in the UIT condition and the proctored verification test condition. The hypothesis of no-change in ability estimates ($H_0: \theta_2 = \theta_1$) between UIT and proctored verification tests condition is tested against an alternative hypothesis ($H_a: \theta_2 > \theta_1$) using a *Z* test. The *Z* test has the following form.

$$Z_{GD} = \frac{|\hat{\theta}_2 - \hat{\theta}_1|}{\sqrt{SE_2^2 + SE_1^2}}, \quad (26)$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates of θ from UIT and a proctored verification test and SE_1^2 and SE_2^2 are squared standard errors associated with $\hat{\theta}_1$ and $\hat{\theta}_2$,

respectively. The standard error associated with $\hat{\theta}$ is an inverse square root of the test information, evaluated at $\widehat{\theta}$. Guo and Drasgow (2010) reasoned that θ_1 and θ_2 are independent due to the property of local independence. θ_1 and θ_2 follow an approximately normal distribution given sufficient test length. Thus, Z_{GD} also follows a standard normal distribution under the null hypothesis. Guo and Drasgow (2010) reported that the Z test demonstrated high power and low Type I error in detecting dishonest applicants. The Type I error as reported by them remained around 0.01 (for $\alpha = 0.01$) when θ was around 0 on the continuum. The Type I error decreased as θ moved away from 0.

Likelihood Ratio Test by Finkleman et al. (2010)

A likelihood ratio (LR) test is a statistical test based on the ratio of two likelihoods: the maximum of a likelihood function over the parameters with restrictions of the null hypothesis and maximum over the larger set of parameters without the restrictions. The LR statistic is defined as (Neyman & Pearson, 1928)

$$\text{LR} = -2\left[\frac{L(\hat{\theta}_0|\mathbf{u})}{L(\hat{\theta}_a|\mathbf{u})}\right] = -2\left[l(\hat{\theta}_0) - l(\hat{\theta}_a)\right], \tag{27}$$

where $\hat{\theta}_0$ is the restricted maximum likelihood estimate under the null hypothesis, $\hat{\theta}_a$ is the unrestricted maximum likelihood estimate under the alternative hypothesis, and $l(\cdot)$ is the logarithm of the likelihood function. The LR statistic approximates a chi-square distribution with 1 degree of freedom under the null hypothesis (Wilks, 1938).

Finkleman et al. (2010) applied the LR statistic in testing the significance of individual change. They defined the LR statistic in testing the null hypothesis of change $H_0: \theta_2 = \theta_1$, as

$$\mathrm{LR}_{FI} = -2\left[\frac{L\left(\hat{\theta}_p \middle| \mathbf{u_{1+2}}\right)}{L\left(\hat{\theta}_1 \middle| \mathbf{u_1}\right) \times L\left(\hat{\theta}_2 \middle| \mathbf{u_2}\right)}\right], \tag{28}$$

where $\mathbf{u}_1$ and $\mathbf{u}_1$ denote response vectors from Occasion 1 and Occasion 2, respectively, and $\mathbf{u_{1+2}}$ is a combined response vector from the two occasions. Under the null hypothesis, $\hat{\theta}_p$ is the value that maximizes the likelihood in the numerator, whereas the likelihood function in the denominator is maximized when estimating the MLEs separately at each occasion, which are denoted as $\hat{\theta}_1$ and $\hat{\theta}_2$. The statistic is compared to a chi-square distribution with 1 degree of freedom to determine the statistical significance of change. Finkleman et al. (2010) and Lee (2015) demonstrated that this approach resulted in desirable Type I error rates and power compared to the confidence interval approach. The LR_{FI} test exhibited more power as well as better Type I error compared to the *Z* test. However, the *Z* approach showed more consistent performance in terms of striking a balance in achieving desirable Type I error and power.

Likelihood Ratio Test by Guo and Drasgow (2010)

The LR statistic in Guo and Drasgow (2010) was based on the likelihoods of the response vectors from the UIT and the proctored verification test,

$$\mathrm{LR}_{GD} = \frac{L(\mathbf{u})L(\mathbf{v})}{L(\mathbf{u},\mathbf{v})}, \tag{29}$$

where $L(\mathbf{u})$, $L(\mathbf{v})$, $L(\mathbf{u},\mathbf{v})$ are the likelihood of observing the responses in the UIT, the proctored verification test, and the two response vectors together, each of which was defined as

$$L(\mathbf{u}) = \int \prod_{i=1}^{n_u} P_i(\theta)^{u_i}[1 - P_i(\theta)]^{1-u_i} f(\theta) d\theta\,, \tag{30}$$

$$L(\mathbf{v}) = \int \prod_{i=1}^{n_v} P_j(\theta)^{v_i}[1 - P_i(\theta)]^{1-v_i} f(\theta) d\theta \,, \tag{31}$$

$$L(\mathbf{u}, \mathbf{v}) = \int \left\{ \prod_{i=1}^{n_u} P_i(\theta)^{u_i}[1 - P_i(\theta)]^{1-u_i} \right\} \left\{ \prod_{i=1}^{n_v} P_i(\theta)^{v_i}[1 - P_i(\theta)]^{1-v_i} \right\} f(\theta) d\theta. \tag{32}$$

The integrations were numerically approximated using θs from −4 to 4 with an 0.1 increment. The critical value was determined using a simulation to have Type I error approximately 0.01 since the distribution of the likelihood ratio in Equation 29 is unknown. The LR_{GD} statistics was evaluated and compared based on observed Type I error and power in detecting suspicious cheating (i.e., decrease in θ in the verification test from UIT).

The methods in Guo and Drasgow (2010) are limited by the fact that the two statistics are not parameterized based on the hypotheses being tested. Instead of defining the statistics with reference to the null hypothesis parameters, the Z_{GD} statistic is defined using MLE in each test. Hence, it is questionable if it really follows a standard normal distribution. Similarly, the LR_{GD} statistic is defined as the ratio between the likelihoods of the observed responses, without taking into account the hypotheses. Both Z_{GD} as well as LR_{GD} showed reasonably good performance for power (0.97 to 1.0) only when there were large discrepancies between the two θ distributions. Finkelman et al.'s (2010) hypothesis testing methods are similar in name to those of Guo and Drasgow (2010) but are more firmly rooted in appropriate statistical theory. Finkleman et al.'s *Z*-test (2010) used the standardized difference in MLE θ estimates and the significance of the test statistic is determined from the standardized normal distribution, and their LR statistic evaluates two likelihood functions under the null and alternative hypotheses of change.

Score Test

Rao (1948) introduced the score test (ST) as an alternative way to use the likelihood function to perform large-sample inference. The ST statistic uses the slope and expected curvature of the log-likelihood function instead of the differences in log-likelihoods. The general form of the ST statistic is defined as

$$\text{ST} = \frac{s(\theta_0|\mathbf{u})^2}{I(\theta_0)}, \tag{33}$$

where $s(\theta_0|\mathbf{u}) = \frac{\delta \ln l(\theta|\mathbf{u})}{d\theta}$, the first derivative of the log-likelihood, which is called a score function, and $I(\theta)$ is test information, both evaluated at θ_0. Since $E[s(\theta|\mathbf{u})] = 0$ and $\text{var}[s(\theta|\mathbf{u})] = I(\theta)$, the ratio of the score function to its null standard error has an approximate standard normal distribution, that is,

$$\frac{s(\theta_0|\mathbf{u})}{\sqrt{I(\theta_0)}} \xrightarrow{D} N(0,1) \tag{34}$$

by the central limit theorem. The ST statistic is the squared value of Equation 34 and follows asymptotically a chi-square distribution with 1 degree of freedom. In testing the hypothesis of no-change, $H_0: \theta_1 = \theta_2$, Lee (2015) defined the score statistic as

$$\text{ST} = \frac{s(\hat{\theta}_p|\mathbf{u_1})^2}{I_1(\hat{\theta}_p)} + \frac{s(\hat{\theta}_p|\mathbf{u_2})^2}{I_2(\hat{\theta}_p)}. \tag{35}$$

In general, the ST statistic is simple to compute as it depends only on estimation of parameters under the null hypothesis, whereas the LR statistic requires estimates both under the null and alternative hypotheses. As the sample size increases to infinity, the ST is asymptotically equivalent to the LR test in the first-order approximation (Rao, 1965; Chandra & Joshi, 1983; Cox & Hinkley, 1974). In finite samples, the two tests tend to

generate somewhat different test statistics. For example, when the model is linear, the LR statistic tends to be greater than or equal to the ST statistic (Johnston & DiNardo, 1977, p.150). The second-order powers of the two tests are also different but neither dominates the other (Taniguchi, 1988, 1991).

Lee (2015) reasoned that if the null hypothesis is true, the MLE at each occasion $\hat{\theta}_1$ and $\hat{\theta}_2$ will be close to $\hat{\theta}_p$ and the slope of the log-likelihood at $\hat{\theta}_p$ for each occasion will also be close to zero. Hence, a smaller ST statistic will be obtained under the null hypothesis. The statistic is then compared to a chi-square distribution with 1 degree of freedom to determine the significance of change. Lee (2015) reported that this test resulted in a desirable Type I error and power in detecting individual change.

Kullback-Leibler Divergence Test

The Kullback-Leibler divergence (KLD) is a measure of distance between two distributions. The use of KLD to detect significance of change was first proposed by Wang (2014) for a conventional test using a multidimensional IRT model and multivariate normal prior. Lee (2015) used the KLD test for AMC with a unidimensional IRT model using both normal (KLD-N) and uniform priors (KLD-U). The KLD (Kullback & Leibler, 1951) between the prior distribution at Occasion 1, $\pi_1(\theta|\mathbf{u_1})$, and the posterior distribution at Occasion 2, $\pi_2(\theta|\mathbf{u_2})$, is defined as

$$\begin{aligned} \mathrm{KLD}(\pi_1 \ \pi_2) &= E_{\pi_1}\left[\ln\frac{\pi_1(\theta|\mathbf{u_1})}{\pi_2(\theta|\mathbf{u_2})}\right] \\ &= \int_{-\infty}^{\infty} \pi_1(\theta|\mathbf{u_1}) \ln\frac{\pi_1(\theta|\mathbf{u_1})}{\pi_2(\theta|\mathbf{u_2})} d\theta. \end{aligned} \tag{36}$$

The value of KLD is always nonnegative and is zero if and only if two distributions are identical. Larger KLD indicates that the two distributions differ and, in the present

application, that change has occurred between the two measurement occasions: if there is no change, KLD will be close to zero. From Equation 36, the KLD statistic can be derived down (see Lee, 2015 for complete derivation) to

$$\mathrm{KLD}(\pi_1 \ \pi_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}. \tag{37}$$

where μ_1 and μ_2 are population means of the distributions of $\pi_1(\theta)$ and $\pi_2(\theta)$, respectively and σ_2 is the standard deviation of the distribution of $\pi_2(\theta)$. Let $Y = \frac{(\mu_1-\mu_1)}{\sqrt{2}\sigma_2} \sim N(0,1)$. Then KLD is distributed as a chi-square with one degree of freedom. For each examinee, the KLD statistic is calculated and compared to the chi-square distribution with 1 degree of freedom to determine the significance of change.

The application of Kullback-Leibler information in CAT item selection was first introduced by Chang and Ying (1996). As described in Equation 36, KL information, or KL divergence, is a general measure for the distance between two distributions. When applied to CAT item selection, the larger value of the KL information indicates that the item better discriminates between two distributions, or equivalently, between the values of the parameters that index them (Lehmann & Casella, 1998). Finkelman et al. (2010) and Lee (2015) applied KL information in the context of detecting individual change by selecting Occasion 2 items that best differentiated θ_2 from θ_1. Finkelman et al. (2010) used $\hat{\theta}_2$ as the best estimate of θ_2 under the alternative hypothesis, and $\hat{\theta}_p$ as the best estimate of $\hat{\theta}_1$ under the null hypothesis in computing KL information. Lee (2015) used a modified version of Finkleman et al.'s (2010) KL test by substituting $\hat{\theta}_p$ by $\hat{\theta}_1$, thereby defining the KL statistic as

$$\mathrm{KLD}\left(\hat{\theta}_1 \;\hat{\theta}_1\right) = E\left[\ln \frac{\hat{\theta}_2|\mathbf{u}_i)}{\pi_2\left(\hat{\theta}_1|\mathbf{u}_i\right)}\right]. \tag{38}$$

In Lee's (2015) study, KLD-U resulted in Type I error of around 0.05 and of around 0.042 for KLD-N. Consequently, KLD-U resulted in higher power than that observed for KLD-N. Observed power for KLD-U was around 0.64 to 0.84 for medium ($\Delta = 1.0$) to large ($\Delta = 1.5$) change, and KLD-N resulted in observed power of around 0.61 to 0.81 for medium and large change, respectively.

Wang and Weiss (2017) extended the research by Finkleman et al. (2010) and Lee (2015) by generalizing the hypothesis tests in the AMC framework to evaluating change on multiple latent traits. They proposed a multivariate *Z* test, a multivariate likelihood ratio test, a multivariate score test, and a Kullback-Leibler test for the two occasion case. Their simulation results showed that the hypothesis tests for the multivariate θ case resulted in low Type I error and high power, showing promising results in simulated as well as real data.

Limitations of the Existing Methods

The hypothesis testing methods for detecting individual change discussed by Finkleman et al. (2010) and Lee (2015) are promising in terms of Type I error and power. However, their methods are defined in terms of two measurement occasions. Hence, they are limited in use. When an examinee is measured on more than two occasions, the hypothesis testing methods discussed above might not be the most suitable approach to measure individual change. For example, when an examinee is measured on three occasions, it becomes necessary to implement the hypothesis test to detect significance of change between all the pairs of occasions, i.e., for three pairs of differences in ability estimates. Such multiple significance testings will inflate the Type I error. This problem

can be overcome by adjusting for the inflated Type I error or by using some kind of omnibus hypothesis test for two or more occasions, instead of multiple testing. Another problem with multiple testing is, as the number of testing occasions grows further, using hypothesis tests designed for two occasions will result in having to calculate several statistics for every examinee, each of which will then be compared to the $1 - \alpha/2^{\text{th}}$ quantile of the appropriate distribution with the adjusted Type I error. Therefore, developing omnibus hypothesis tests to detect individual change seems like a more appropriate alternative to multiple hypothesis tests. Developing such omnibus tests and investigating performance of the hypothesis tests in case of multiple occasions is a very logical and vital extension of the previous research.

The present study expands the work by Lee (2015) and Finkleman et al. (2010) and proposes and evaluates the performance of new omnibus hypothesis tests which are derived from hypothesis tests based on two occasions. The proposed tests are generalized omnibus tests and can be used for multiple testing occasions ($t \geq 2$).

Chapter 2: Method

This study involved conceptualizing and testing the performance of omnibus hypothesis tests for detecting change at an individual level in the context of IRT and CAT. The simulation design was similar to that used by Finkleman et al. (2010) and Lee (2015). However, the present study expanded their research to incorporate more testing conditions and implemented change over multiple occasions instead of two occasions. Amount of change was varied in different combinations over multiple occasions, resulting in various change patterns. New omnibus hypothesis tests for multiple testing occasions were proposed and their performance in terms of Type I error and power was evaluated.

New Omnibus Hypothesis Tests

Generalization of Z tests

Finkleman, Weiss, and Kim-Kang (2010) described a *Z* test to determine if psychometrically significant change occurred within an examinee between two occasions. The *Z* statistic presented in Equation 25 is distributed $Z_{FI} \sim \mathrm{N}(0,1)$. Guo and Drasgow also described a *Z* test to determine if scores obtained from a person differ significantly between two testing occasions. Their *Z* statistic as presented in Equation 26 is also assumed to be distributed $Z_{GD} \sim \mathrm{N}(0,1)$.

When an examinee is measured at three occasions, three different Z_{FI} or Z_{GD} statistics can be obtained between Occasion 1 – Occasion 2 (Z_{12}), Occasion 2 – Occasion 3 (Z_{23}) and Occasion 1 – Occasion 3 (Z_{13}). To generalize, if an examinee is measured at *t* occasions, $t(t-1)/2 = k$ unique comparisons can be made between the ability estimates, and *k Z* statistics can be obtained between the occasions. Since *k* independent and standard

normal variables constitute a χ^2 distribution on k degrees of freedom, Finkleman's chi-square test can be constituted from Finkleman, Weiss, and Kim-Kang's Z test (2010) as

$$\chi^2_{FI} = \sum_{i=1}^{k} Z^2_{i(FI)} \tag{39}$$

and Guo and Drasgow's chi-square test can be constituted from their *Z* test as

$$\chi^2_{GD} = \sum_{i=1}^{k} Z^2_{i(GD)}, \tag{40}$$

where Z_i is Finkleman's or Guo and Drasgow's Z statistic obtained between any two occasions and $\sum_{i=1}^{k} Z_i^2$ is approximately χ^2 distributed with k degrees of freedom. Observed change can be determined as significant when the χ^2_{FI} or χ^2_{GD} statistic exceeds the $1-\alpha$ quantile of the chi-square distribution with k degrees of freedom.

It must be noted that both χ^2_{FI} and χ^2_{GD} are approximate χ^2 distributions on k degrees of freedom as the *Z*s are not strictly independent. They are obtained on the basis of within-person scores at different occasions. However, previous research (Wang & Weiss, 2017) shows that these approaches work well in the case of multidimensional IRT to have reasonable Type I error and power. Their Type I error and power were evaluated in the present study.

Analysis of Variance

The current study also investigated an analysis of variance (ANOVA) framework to test the hypothesis of no-change within an examinee. Two kinds of *F* ratio statistics, differing in their formalization, were proposed and investigated. For measuring change within person in a psychometric framework, an *F* ratio can be defined as

$$\mathrm{F1}_{T-1,N-T} = \frac{[\sum_{t=1}^{T}(\hat{\theta}_t - \hat{\theta}_G)^2 / n_t]/T - 1}{\sum_{t=1}^{T} I_t(\hat{\theta}_t)^{-1}/N - T}, \quad (41)$$

where, T = total number of testing occasions, $\hat{\theta}_t$ = estimated θ at t^{th} occasion, $\hat{\theta}_G$ = grand mean of $\hat{\theta}$s obtained over all occasions, $I_t(\hat{\theta}_t)$ = test information obtained at t^{th} $\hat{\theta}$, and N = $\sum_{t=1}^{T} n_t$ = total number of items used across all tests and occasions. It should be noted that for a fixed-length CAT, n_t would be the same at all occasions. However, for a variable length CAT, n_t may vary across occasions.

The numerator of the F1 statistic represents the "between sum of squares" in the psychometric framework divided by $T - 1$ degrees of freedom. In the standard F test of statistical significance, the sum of squared variation of group means around the grand mean is scaled up by the number of observations in each group. However, in the psychometric framework of individual change, only one observation is obtained at each occasion. Therefore, one way to formulate the F statistic is by scaling down the sum of squared variation of $\hat{\theta}_t$ around $\hat{\theta}_G$, by dividing it by n_t in the numerator, as the items are assumed to constitute $\hat{\theta}_t$. Thus, the variation is scaled down because of lack of multiple observations. The denominator of Equation 41 represents the within sum of squares (from a psychometric error framework) divided by $N - T$ degrees of freedom. The sum of reciprocals of test information evaluated at $\hat{\theta}_t$ is assumed to account for variability attributed to factors unrelated to individual change.

Assuming an individual is measured at three occasions using fixed-length tests, Equation 41 can be expanded as

$$\mathrm{F1}_{3-1,N-3} = \frac{([(\hat{\theta}_1 - \hat{\theta}_G)^2 + (\hat{\theta}_2 - \hat{\theta}_G)^2 + (\hat{\theta}_3 - \hat{\theta}_G)^2]/n_t])/(3-1)}{[I_1(\hat{\theta}_1)^{-1} + I_2(\hat{\theta}_2)^{-1} + I_3(\hat{\theta}_3)^{-1}]/N - 3}. \quad (42)$$

Observed change can be determined as significant when the *F* statistic exceeds the $1 - \alpha$ quantile of the *F* distribution on $T - 1$ degrees of freedom for the numerator and $N - T$ degrees of freedom for the denominator.

The current study also proposes another form of *F* statistic, slightly different in its conceptualization defined as

$$\mathrm{F2}_{T-1,N-T} = \frac{T \times \mathrm{var}(\hat{\theta}_t)}{\sum_{t=1}^{T} I_t(\hat{\theta}_t)^{-1}}, \quad (43)$$

In Equation 43 variation of $\hat{\theta}_t$ is scaled up to the total number of occasions (*T*), instead of scaling down by the total number of items as in Equation 42. The difference between the two forms of *F* statistics is inclusion of the degrees of freedom in the formula. In Equation 41, the psychometric equivalent of "between sum of squares" is divided by between degrees of freedom and the psychometric equivalent of "within sum of squares" is divided by within degrees of freedom. However, in Equation 43, degrees of freedom do not appear in the F2 formula.

In F1 as well as in F2, error is assumed to reside in the items used to measure θ, unlike the *F* test used to measure statistical significance. In the latter, error is attributed to individual differences. The "mean square within" reflects variation due to individual variability. In the psychometric framework of measuring individual change, items which are used to measure the individual's ability vary in their characteristics in capturing that ability/latent trait. Hence, it is assumed that error is inherent in the items and variability due to other factors can be attributed to the reciprocal of the test information.

Likelihood Ratio Test

The extension of Finkleman et al.'s LR test to multiple occasion is straightforward. In the context of two occasions, Finkleman et al. (2010) defined a likelihood ratio (LR) statistic as presented in Equation 28 to determine significance of individual change. When there are three or more occasions, the same LR statistic can be expressed as

$$\text{LR} = -2\ln\frac{L\left(\hat{\theta}_p\middle|\mathbf{u}_{1+2+\cdots T}\right)}{L\left(\hat{\theta}_1\middle|\mathbf{u}_1\right)\times L\left(\hat{\theta}_2\middle|\mathbf{u}_2\right)\times \ldots \times L\left(\hat{\theta}_T\middle|\mathbf{u}_T\right)}. \tag{44}$$

where $\mathbf{u}_{t=1,2,\ldots,T}$ is a combined response vector from the all the measurement occasions. Observed change can be determined as significant when the LR statistic exceeds the $1-\alpha$ quantile of the chi-square distribution with $T-1$ degrees of freedom. The simplified version of Equation 44 for three measurement occasions can be re-written as

$$\text{LR} = -2\ln\frac{L\left(\hat{\theta}_p\middle|\mathbf{u}_{1+2+3}\right)}{L\left(\hat{\theta}_1\middle|\mathbf{u}_1\right)\times L\left(\hat{\theta}_2\middle|\mathbf{u}_2\right)\times L\left(\hat{\theta}_3\middle|\mathbf{u}_3\right)}. \tag{45}$$

Score Test

Extension of the ST from the two occasion case to the multi-occasion case follows directly from Equation 35. Lee (2015) used the ST to determine the significance of change in the case of two occasions as presented in Equation 35. The same test can be extended when an individual is measured at more than two occasions to determine whether there is significant change at any of the occasions. For multiple occasions, this extension can be presented in the following form

$$\text{ST} = \frac{s(\hat{\theta}_p|\mathbf{u_1})^2}{I_1(\hat{\theta}_p)} + \frac{s(\hat{\theta}_p|\mathbf{u_2})^2}{I_2(\hat{\theta}_p)} + \cdots + \frac{s(\hat{\theta}_p|\mathbf{u}_T)^2}{I_T(\hat{\theta}_p)}. \tag{46}$$

Observed change can be determined as significant when the ST statistic exceeds the $1-\alpha$ quantile of the chi-square distribution with $T-1$ degrees of freedom. Thus, for example, for the three-occasion cases, the ST statistic can be simplified to

$$\mathrm{ST} = \frac{s(\hat{\theta}_p|\mathbf{u_1})^2}{I_1(\hat{\theta}_p)} + \frac{s(\hat{\theta}_p|\mathbf{u_2})^2}{I_2(\hat{\theta}_p)} + \frac{s(\hat{\theta}_p|\mathbf{u_3})^2}{I_3(\hat{\theta}_p)}. \quad (47)$$

Simulation Design

Like Lee (2015) and Finkleman et al. (2010), nine different θ levels (–2, –1.5, –1, –0.5, 0, 0.5, 1, 1.5, 2) were used as baseline levels from which change was simulated. The amount of change was varied in four levels of no-change ($\Delta = 0$), small change ($\Delta = 0.5$ SDs), medium change ($\Delta = 1.0$ SDs) and large change ($\Delta = 1.5$ SDs) on the θ scale with mean 0 and SD = 1.

Measurement Occasions and Patterns of Change

Three measurement occasions were used to implement and assess change. Change in varying amounts ($\Delta = 0$, 0.5, 1.0, 1.5 SDs) was introduced after Occasion 1 and Occasion 2. 10 unique combinations were formed between the occasions and the level of change. Table 2.1 presents these 10 unique patterns of change. Each level of amount of change between Occasion 1 and Occasion 2 was crossed with other levels of change between Occasion 2 and Occasion 3. Out of the 10 change patterns presented in Table 2.1, the patterns in which amount of change remained the same between the occasions are *linear change patterns*. For example, $\Delta = 0.5$ between Occasion 1 and Occasion 2 and $\Delta = 0.5$ between Occasion 2 and Occasion 3 is a linear pattern of change (L1). Similarly, $\Delta = 1.0$, 1.0 (L2) and $\Delta = 1.5$, 1.5 (L3) are also linear patterns of change. The patterns in which the amount of change varied between the occasions are *non-linear patterns of change*. Thus,

$\Delta = 0, 0.5$ (NL1), $\Delta = 0, 1.0$ (NL2), $\Delta = 0, 1.5$ (NL3), $\Delta = 0.5, 1.0$ (NL4), $\Delta = 1.0, 1.5$ (NL5), and $\Delta = 1.0, 1.5$ (NL6) are non-linear patterns of change. $\Delta = 0, 0$ represents the *no-change* condition. Only the unique patterns of change between the occasions were considered, as the goal of this research was test the performance of omnibus hypothesis testing methods in detecting true change.

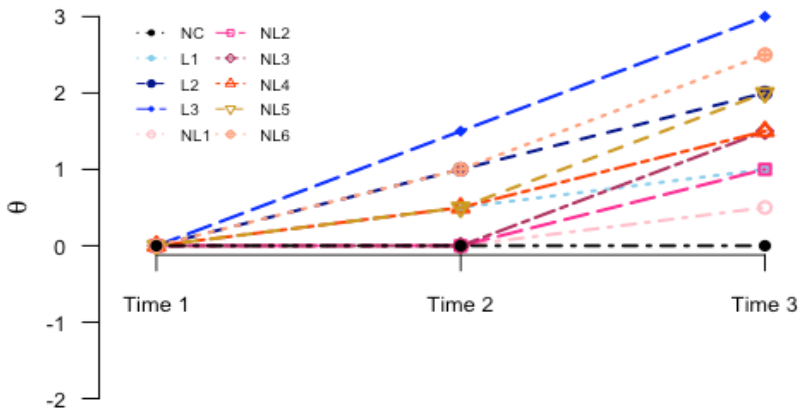
Table 2.1: Unique Combinations of Amount of Change Crossed with Occasions

Occasion 1	Occasion 2	Occasion 3
$\Delta = 0.0$		$\Delta = 0.0$ $\Delta = 0.5$ $\Delta = 1.0$ $\Delta = 1.5$
$\Delta = 0.5$		$\Delta = 0.5$ $\Delta = 1.0$ $\Delta = 1.5$
$\Delta = 1.0$		$\Delta = 1.0$ $\Delta = 1.5$
$\Delta = 1.5$		$\Delta = 1.5$

The upper limit on θ after introducing change was set to $\theta = 3$. All the 10 change patterns between the occasions and amount of change were crossed with $\theta = -2$, $\theta = -1.5$, $\theta = -1$, $\theta = -0.5$, and $\theta = 0$. $\theta = 0.5$ was crossed with nine change patterns (except $\Delta = 1.5, 1.5$) to keep the upper limit at $\theta = 3$. Similarly, $\theta = 1.0$ was crossed with eight change patterns (except $\Delta = 1.5, 1.5$ and $\Delta = 1.0, 1.5$), $\theta = 1.5$ was crossed with six change patterns (except $\Delta = 1.5, 1.5$, $\Delta = 1.0, 1.5$, $\Delta = 1.0, 1.0$, and $\Delta = 0.5, 1.5$) and $\theta = 2.0$ was crossed with four unique change patterns (except $\Delta = 1.5, 1.5$, $\Delta = 1.0, 1.5$, $\Delta = 1.0, 1.0$, $\Delta = 0.5,$

1.5, Δ = 0.5, 1.0, and Δ = 0.0, 1.5). Thus, there were 77 total combinations between θ levels and the 10 change patterns. Figure 2.1 presents the 10 change patterns at $\theta = 0$. The no-change (NC) condition is represented by the black dashed line. Three linear change patterns (L1, L2 and L3) are presented by straight lines in which the amount of change remained consistent across the occasions. Six non-linear change patterns (NL1, NL2, NL3, NL4, NL5 and NL6) are presented by inclined lines in which the amount of change varied between the occasions.

Figure 2.1: Change Patterns at $\theta = 0$



Item Banks

30-item fixed-length CATs with varying item discriminations and peakedness were used. Table 2.2 summarizes different parameters used for creating six different types of item banks, varying in discrimination and difficulty.

The high discrimination (HD) bank was created by generating the item discrimination parameter (a_i) from a normal distribution with a mean of 1.5 and standard deviation of 0.15 [$a \sim \mathrm{N}(1.5, 0.15)$]. The medium discrimination (MD) bank was created by generating the item discrimination parameter from a normal distribution with a mean of 1.0 [$a \sim \mathrm{N}(1.0, 0.15),$], and the low discrimination (LD) bank was created by generating

the item discrimination parameter from a normal distribution with a mean of 0.6 $[a \sim \mathrm{N}(0.6, 0.15)]$.

Table 2.2: Parameters for Varying Item Bank Conditions

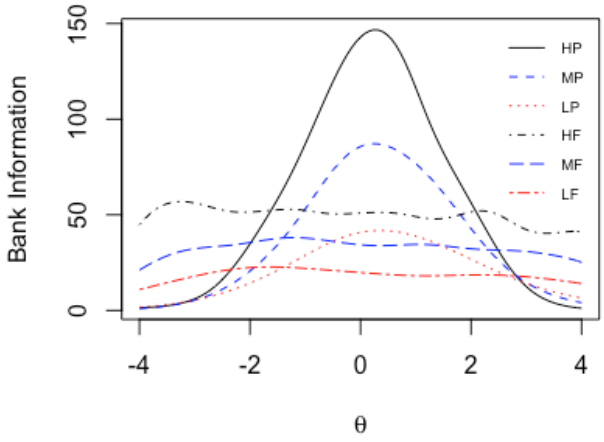
	Discrimination			Difficulty	
	High	Medium	Low	Flat	Peaked
a_i	~ N(1.5, 0.15)	~ N(1.0, 0.15)	~ N(0.6, 0.15)	–	
b_i	–			~ U(– 4.5, 4.5)	~ U(0, 0.8)
c_i	0.2				

Two sets of difficulty parameters (b_i) were used to represent flat or peaked CATs. One set of difficulty parameters was generated from a uniform distribution $[b \sim \mathrm{U}(-4.5, 4.5)]$ to create a flat item bank (FB) and another set was generated from a normal distribution $[b \sim \mathrm{N}(0.0, 0.8)]$ to represent a more realistic peaked item bank (PB). Three discrimination conditions crossed with two information types resulted in six types of item banks – High Flat (HF), High Peaked (HP), Medium Flat (MF), Medium Peaked (MP), Low Flat (LF) and Low Peaked (LP). Each of these six combinations (HF, HP, MF, MP, LF, LP) of item banks consisted of 300 items. The lower asymptote (c_i) was kept constant at 0.2. Six different bank information functions are depicted in Figure 2.2.

Data Generation and Scoring

Item responses of 1,000 examinees at each of the θ levels were generated for all the conditions in accordance with the three-parameter logistic IRT model (Birnbaum, 1968), defined in Equation 17. Item responses were generated using a monte-carlo

Figure 2.2: Test Information Functions of Six CAT Item Banks



simulation, which involves generating a random uniform number from U [0,1] for each item-person interaction. The item response was treated as correct and coded as 1 if the probability of answering the item correctly was greater than the randomly generated number. The item response was treated as incorrect and coded as 0 if the probability of answering the item correctly was less than the randomly generated number. θs were estimated using maximum likelihood estimation (MLE) with the bounds of $[-4, 4]$, with a temporary use of expected a posteriori (EAP) for non-mixed response patterns.

AMC Procedures

The initial θ was set to zero for all examinees for Occasion 1. The final maximum likelihood (ML) estimate of θ at Occasion 1 ($\hat{\theta}_1$) was used as the starting θ estimate for the Occasion 2 CAT and the final ML estimate of θ at Occasion 2 ($\hat{\theta}_2$) was used as the starting θ estimate for the Occasion 3 CAT. At Occasion 1, 2 and 3, items were selected based on Fisher information at the current θ estimate. Six omnibus hypothesis testing statistics for multiple occasions – χ^2_{FI}(Equation 39), χ^2_{GD}(Equation 40), F1(Equation 41), F2 (Equation 43), LR (Equation 44), and ST (Equation 46), – were computed for each

examinee at the final stage of Occasion 3. The test was terminated after administration of 30 items. The same test length was used across three occasions.

Conditions

This study used 77 (3 occasions crossed with amount of change combinations at 9 θ levels) response conditions and 3 (item discrimination) × 2 (peakedness) test conditions. Item responses were generated independently for all 77 response conditions. Six hypothesis testing methods were fully crossed with all of the 77 response conditions. Thus, the design was a 77 (3 occasions crossed with amount of change combinations at 9 θ levels) × 3 (item discrimination) × 2 (peakedness) × 6 (hypothesis testing method) ANOVA design, resulting in 2,772 total conditions. All simulations and procedures were performed in R (R Core Team, 2016).

Dependent Variables

Type I Error and Power

The performance of the six hypothesis tests in the detection of true change at an individual level was evaluated in terms of Type I error and power. Type I error was determined by the proportion of times the hypothesis of no-change was rejected under the no-change condition. The proportion of times the hypothesis of no-change was rejected under the conditions of small, medium and large change conditions (and their sub-conditions of different change patterns) was operationalized as power. Calculation of Type I error and power are described in more detail in the "Replications" section below.

Agreement Between Methods

Agreement between methods was evaluated in terms of proportion of times they agreed in detecting significant change across each set of 1,000 simulees. Estimation of

Type I error, power, and agreement between the methods was made by averaging the proportions across the replications (as described below) in each condition.

Effect Size

An ANOVA design was used to summarize the main effects and interactions on each dependent variable. Effect size (η^2) was computed for each effect in ANOVA. η^2 is defined as a ratio of sums of squares,

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}. \tag{48}$$

Type I error rate was the dependent variable in the no-change condition, and power, i.e., proportion of correct classifications, was used as the dependent variable in the remaining conditions of change. η^2 values were multiplied by 100 to express them as percentages.

Replications

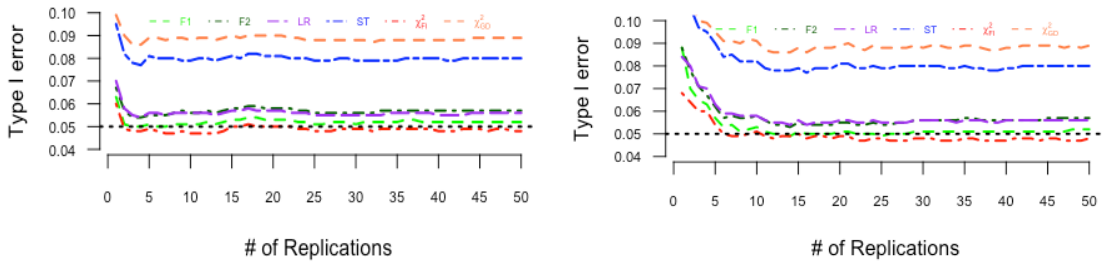
Pilot studies were conducted on a few different conditions to determine the number of replications required in order to ensure stability of results. These conditions were selected to serve a representational basis for running the replications across other conditions. Thus, within discrimination, high and low discrimination conditions crossed with flat and peak information were selected. All hypothesis tests were run across high discrimination/flat (HF) and low discrimination/peaked (LP) as pilot replication conditions. 1 to 50 replications were run for no-change, L1 and NL3 change patterns using HF and low LP item banks at θ = 0. Item responses of 250 or 1,000 simulees were generated under each condition. Mean Type I error and power were plotted as a function of number of replications to determine the number of replications at which the results stabilized across the pilot conditions. On a representational basis, the results of the replications are presented

in Figures 2.3 and 2.4 for two conditions. However, they were generally consistent across conditions.

Figure 2.3: Mean Type I Error Conditional on Replications for HF and LP Item Banks

a. 50 replications/1000 simulees in HF

b.50 replications/250 simulees in HF



c. 50 replications/1000 simulees in LP

d. 50 replications/250 simulees in LP

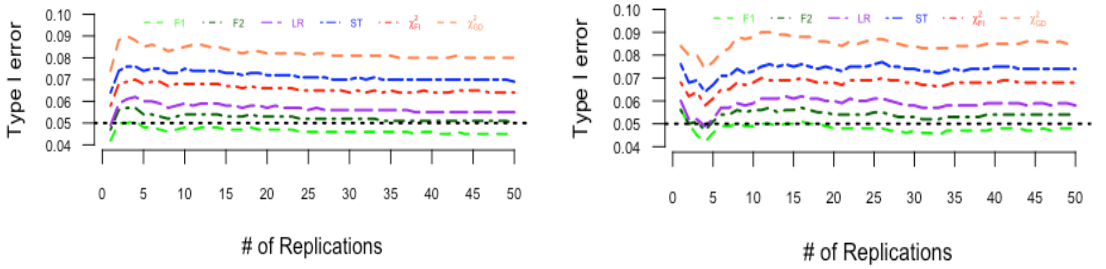
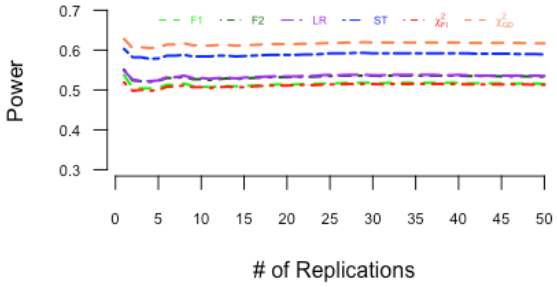


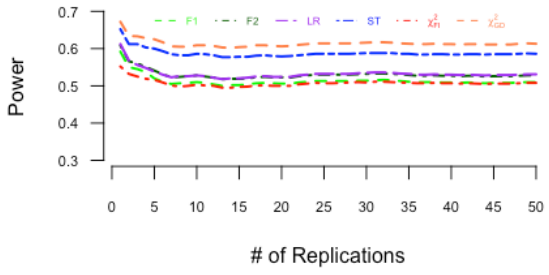
Figure 2.3 presents mean Type I error plotted against replications for two item banks and two numbers of simulees combinations. Figures 2.3a and 2.3b display mean Type I error conditional on number of replications for the high discrimination/flat (HF) item bank with 1,000 and 250 simuless per replication, respectively. Figures 2.3c and 2.3d present Type I error for the low discrimination/peaked (LP) item bank (LP). Type I error stabilized after five replications in both conditions for the LP bank. In Figures 2.3b and 2.3d based on 1,000 simulees, Type I error was consistent after around 15 replications.

Figure 2.4: Mean Power Conditional on Replications for L1 Change Pattern for HF and LP Item Banks

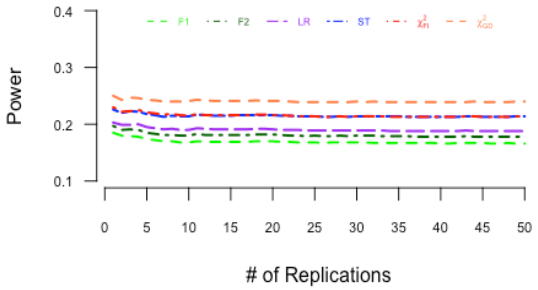
a. 50 replications/1000 simulees in HF-L1



b. 50 replications/250 simulees in HF-L1



c. 50 replications/1000 simulees in LP-L1



d. 50 replications/250 simulees in LP-L1

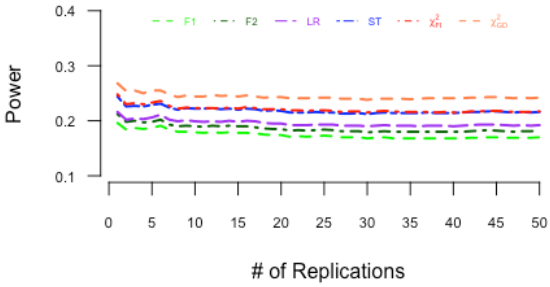


Figure 2.4 depicts power conditional on number of replications for HF and LP item banks for the L1 change pattern. Here. too, when 1,000 simulees were used per replication, results stabilized after 5 replications. When 250 examinees were used per replication, it took around 10 replications for the results to be consistent across replications. Based on these results, each condition was replicated 10 times by generating a new set of 1,000 simulees for each replication. Because the hypothesis tests for detecting change in this study were implemented for each single simulee, this was equivalent to generation 10,000 simulees per condition to compute Type I error and power.

Chapter 3: Results

Type I error

Tables 3.1 through 3.10 present the ANOVA results. The tables show source of variation, sum of squares, degrees of freedom, and η^2 as a percentage. Table 3.1a summarizes the results of a four factor ANOVA on Type I error, through three-way interactions. The largest source of variation accounted for was by the type of statistic (η^2 = 80.93%). Other main effects (namely θ, Discrimination and Information/Peakedness) accounted for less than 5% of total variability. All the interactions accounted for about 5% variation and the error variation was less than 1%.

Table 3.1b presents means and standard deviations of Type I error conditional on different types of statistic. Mean Type I error was around 0.05 for F1, F2, LR and χ^2_{FI} statistics. ST and χ^2_{GD} had high mean Type I error. The standard deviation remained in the range of 0.002 to 0.019 for all the statistics.

Power: Linear Change

Table 3.2a presents results of ANOVA on power for the L1 change pattern, in which amount of change was Δ = 0.5, 0.5 across the three occasions. The maximum proportion of variation was accounted for by Discrimination (90.83%) followed by θ (3.32%). The error variation was under 5%. There were no interactions that accounted for more than 1.54% of the variance.

The differences in the mean power and standard deviation conditional on discrimination conditions are shown in Table 3.2b. The high discrimination condition (HD) resulted in the highest power, followed by medium discrimination (MD) and lastly by low

discrimination (LD). The standard deviation for different conditions varied in the range of 0.031 to 0.053.

Table 3.1a: Results of ANOVA with 3-Way Interaction on Type I Error

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.000735	8	0.91%
Discrimination	0.000386	2	0.48%
Information/Peakedness	0.001040	1	1.28%
Statistic	0.065529	5	**80.93%**
θ × Discrimination	0.000882	16	1.09%
θ × Information	0.000739	8	0.91%
θ × Statistic	0.002864	40	3.54%
Discrimination × Information	0.000013	2	0.02%
Discrimination × Statistic	0.001575	10	1.94%
Information × Statistic	0.001606	5	1.98%
θ × Discrimination × Information	0.000728	16	0.90%
θ × Discrimination × Statistic	0.000773	80	0.95%
θ × Information × Statistic	0.003179	40	3.93%
Discrimination × Information × Statistic	0.000185	10	0.23%
Residuals	0.000740	539,756	0.91%
Total	0.080974	539,999	100.00%

Table 3.1b: Mean and SD of Type I Error Conditional on Statistic

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.053	0.058	0.053	0.082	0.055	0.090
SD	0.002	0.002	0.019	0.006	0.005	0.003

Table 3.2a: Results of ANOVA with 2-Way Interactions on Power for L1 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.5207	8	3.32%
Discrimination	14.2473	2	**90.83%**
Information/Peakedness	0.0699	1	0.45%
Statistic	0.2988	5	1.90%
$\theta \times$ Discrimination	0.0603	16	0.38%
$\theta \times$ Information	0.2418	8	1.54%
$\theta \times$ Statistic	0.0489	40	0.31%
Discrimination × Information	0.0116	2	0.07%
Discrimination × Statistic	0.0473	10	0.30%
Information × Statistic	0.0255	5	0.16%
Residuals	0.114	539,902	0.73%
Total	15.6861	539,999	100%

Table 3.2b: Mean and SD of Type I Error Conditional on Discrimination for L1 Change Pattern

	HD	MD	LD
Mean	0.892	0.660	0.381
SD	0.053	0.064	0.031

Table 3.3a shows ANOVA results on power for the L2 pattern of change. This change pattern was based on a change of Δ = 1.0, 1.0 across three occasions. Discrimination was observed to account for maximum variation with η^2 = 74.12% followed by type of Statistic with η^2 = 5.72%. The amount of variation accounted for by the interactions as well as error was less than 5%.

Table 3.3a: Results of ANOVA with 2-Way Interaction on Power for L2 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.004515	6	1.10%
Discrimination	0.303087	2	**74.12%**
Information/Peakedness	0.010452	1	2.56%
Statistic	0.023379	5	**5.72%**
θ × Discrimination	0.010294	12	2.52%
θ × Information	0.004386	6	1.07%
θ × Statistic	0.007031	30	1.72%
Discrimination × Information	0.017638	2	4.31%
Discrimination × Statistic	0.007707	10	1.88%
Information × Statistic	0.002274	5	0.56%
Residuals	0.018142	419,920	4.44%
Total	0.408905	419,999	100%

Table 3.3b: Mean and SD of Power Conditional on Discrimination for L2 Change Pattern

	HD	MD	LD
Mean	0.996	0.995	0.922
SD	0.004	0.002	0.014

Table 3.3c: Mean and SD of Power Conditional on Statistic for L2 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.969	0.972	0.973	0.978	0.951	0.981
SD	0.007	0.007	0.005	0.006	0.012	0.005

Table 3.3b and 3.3c display means and standard deviations conditional on Discrimination and Statistic, respectively for the L2 change pattern. The high discrimination condition resulted in highest mean power followed by medium and then by low discrimination conditions. Mean power conditional on Statistic varied in the range of 0.951 to 0.981, while the standard deviation varied in the range 0.005 to 0.012. χ^2_{GD}

followed by the ST had the highest mean power, but both also had Type I error rates that deviated most from the expected .05 rate (Table 3.1b).

ANOVA results based on the power observed under L3 linear change pattern as a dependent variable are depicted in Table 3.4a. This change pattern consisted of change of $\Delta = 1.5, 1.5$ across three occasions.

Table 3.4a: Results of ANOVA with 3-Way Interaction on Power for L3 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.002315	4	**5.18%**
Discrimination	0.0003293	2	0.74%
Information/Peakedness	0.0000047	1	0.01%
Statistic	0.0126326	5	**28.28%**
θ × Discrimination	0.0016491	8	3.69%
θ × Information	0.0004231	4	0.95%
θ × Statistic	0.0119652	20	**26.78%**
Discrimination × Information	0.0002874	2	0.64%
Discrimination × Statistic	0.0025763	10	5.77%
Information × Statistic	0.000051	5	0.11%
θ × Discrimination × Information	0.0002343	8	0.52%
θ × Discrimination × Statistic	0.0081308	40	**18.20%**
θ × Information × Statistic	0.0019678	20	4.40%
Discrimination × Information × Statistic	0.0008616	10	1.93%
Residuals	0.0012483	359,860	2.79%
Total	0.0446765	359,999	100.00%

Table 3.4b: Mean and SD of Power Conditional on Statistic for L3 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.999	0.999	1.0	1.0	0.977	1.0
SD	0.0003	0.0003	0.0002	0.0002	0.024	0.0001

Among the main effects, the maximum amount of variation was accounted for by type of statistic (η^2 = 28.28%) followed by the main effect of θ (η^2 = 5.18%). The two-way θ × Statistic interaction accounted for 26.78% of variation and the Discrimination × Statistic interaction accounted for 5.77%. The three-way θ × Discrimination × Statistic interaction accounted for 18.20% of variation. The remaining interactions and error accounted for less than 5% variation. For the main effect of θ, mean power varied from 0.989 to 0.999 (Appendix, Table A2).

Table 3.4b presents means and standard deviations of power conditional on Statistic for the L3 change pattern. For the Statistic conditions, mean power varied from 0.977 to 1.0 and standard deviations from 0.0002 to 0.024.

Means and standard deviations of power conditional on θ and Statistic, a 2-way interaction found significant for L3, can be found in Appendix Table A3. The table shows θ levels up to 0.0, as the L3 change pattern was introduced up to θ = 0.0, and not at higher levels to limit the upper bound at 3.0. For all statistics except χ^2_{FI}, power was very high across the θ range. Mean power for χ^2_{FI} decreased as θ deviated from 0. Means of another 3-way interaction found significant in L3 can be found in Appendix Table A4.

The significant 2 and 3-way interactions are presented in Figures 3.1 and 3.2, respectively. Figure 3.1 shows that all the statistics resulted in very similar mean power across the θ range, except χ^2_{FI} which resulted in power of around 0.94 against that of about 1.0 of other statistics. A 3-way θ × Discrimination × Statistic significant interaction also depicts a similar trend for χ^2_{FI}, with the more prominent differences in power between χ^2_{FI} and other statistics in high and medium discrimination compared to low discrimination conditions.

Figure 3.1: 2-Way θ × Statistic Interaction for the L3 Change Pattern

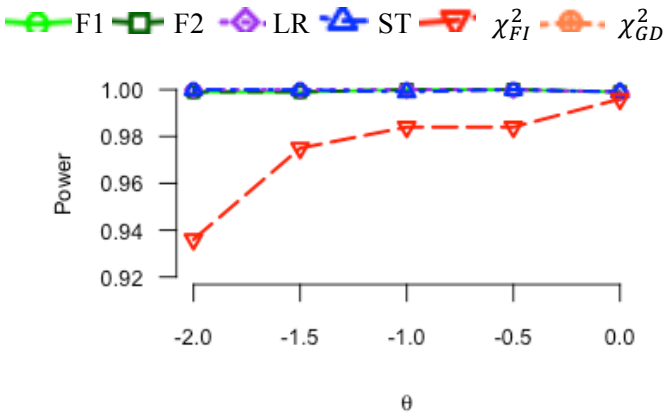
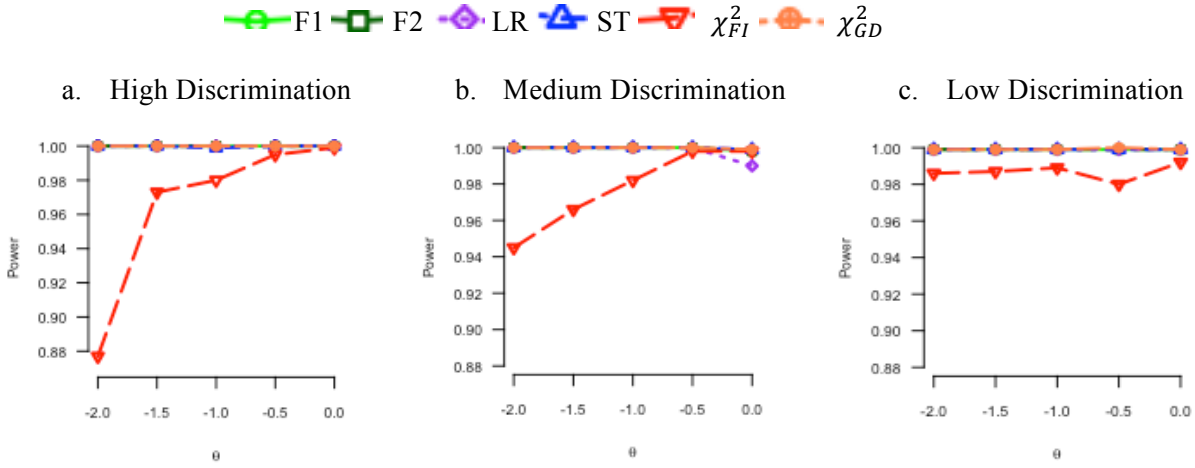


Figure 3.2: 3-Way θ × Discrimination × Statistic Interaction for the L3 Change Pattern



Power: Non-linear Change

Table 3.5 presents ANOVA results for power under non-linear change pattern NL1 (Δ = 0, 0.5 across three occasions). Maximum variation was accounted for by the main effect of Discrimination (84.91%) followed by main effect of Statistic (6.44%). All other main effects, two-way interactions, and error accounted for less than 5% variation.

Means and standard deviation of the significant main effects of Discrimination and Statistic for NL1 are presented in Table 3.5b and 3.5c, respectively. Mean power was highest for the high discrimination condition followed by medium and low discrimination. Standard deviation varied in the range of 0.009 to 0.045. Mean power was highest (0.347)

Table 3.5a: Results of ANOVA with 2-Way Interaction on Power for NL1 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.1524	8	2.48%
Discrimination	5.2078	2	**84.91%**
Information/Peakedness	0.055	1	0.90%
Statistic	0.3949	5	**6.44%**
θ × Discrimination	0.0869	16	1.42%
θ × Information	0.065	8	1.06%
θ × Statistic	0.0279	40	0.45%
Discrimination × Information	0.0235	2	0.38%
Discrimination × Statistic	0.0421	10	0.69%
Information × Statistic	0.0088	5	0.14%
Residuals	0.069	539,902	1.13%
Total	0.0446765	539,999	100%

Table 3.5b: Mean and SD of Power Conditional on Discrimination for NL1 Change Pattern

	HD	MD	LD
Mean	0.472	0.285	0.164
SD	0.045	0.020	0.009

Table 3.5c: Mean and SD of Power Conditional on Statistic for NL1 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.277	0.291	0.290	0.347	0.275	0.364
SD	0.025	0.025	0.021	0.029	0.021	0.027

for ST followed by χ^2_{GD} (0.364) (which again, also had the highest Type I error rates). Mean power for other types of statistic varied in the range of 0.275 to 0.291. Of the other statistics that had adequate control of Type I error rates, LR and F2 had the highest power. Standard deviations conditional on statistic varied in the range of 0.021 to 0.029.

ANOVA results for observed power under the NL2 change pattern (Δ = 0, 1.0 across three occasions) are displayed in Table 3.6a. Similar to NL1, the main effect of Discrimination accounted for maximum variation with η^2 = 91.02%. The variation accounted for by other main effects, two-way interactions and error was less than 5%.

Table 3.6a: Results of ANOVA with 2-Way Interactions on Power for NL2 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.3407	8	2.53%
Discrimination	12.2424	2	**91.02%**
Information/Peakedness	0.0624	1	0.46%
Statistic	0.2974	5	2.21%
θ × Discrimination	0.0363	16	0.27%
θ × Information	0.1683	8	1.25%
θ × Statistic	0.0537	40	0.40%
Discrimination × Information	0.0179	2	0.13%
Discrimination × Statistic	0.0873	10	0.65%
Information × Statistic	0.026	5	0.19%
Residuals	0.1172	539,902	0.87%
Total	13.4496	539,999	100.00%

Table 3.6b: Mean and SD of Power Conditional on Discrimination for NL2 Change Pattern

	HD	MD	LD
Mean	0.948	0.795	0.481
SD	0.030	0.042	0.035

Mean and standard deviation of power for the significant main effect of Discrimination are displayed in Table 3.6b. Mean power for high discrimination was 0.948, that for medium discrimination was 0.79,5 and that for the low discrimination condition

was 0.481. Standard deviations for the discrimination conditions for the NL2 pattern varied in the range of 0.030 to 0.042.

Results of ANOVA on power under NL3 (Δ = 0, 1.5) are shown in Table 3.7a. η^2 was highest for Discrimination (70.58%) followed by that for the effect of Statistic (6.90%). Variation accounted for by other main effects, three-way interactions, and error was less than 5%.

Table 3.7a: Results of ANOVA With 3-Way Interactions on Power for NL3 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.02923	7	1.39%
Discrimination	1.48434	2	**70.58%**
Information/Peakedness	0.02657	1	1.26%
Statistic	0.14515	5	**6.90%**
θ × Discrimination	0.05119	14	2.43%
θ × Information	0.02468	7	1.17%
θ × Statistic	0.08815	35	4.19%
Discrimination × Information	0.03406	2	1.62%
Discrimination × Statistic	0.04262	10	2.03%
Information × Statistic	0.00411	5	0.20%
θ × Discrimination × Information	0.0409	14	1.94%
θ × Discrimination × Statistic	0.07955	70	3.78%
θ × Information × Statistic	0.02728	35	1.30%
Discrimination × Information × Statistic	0.01061	10	0.50%
Residuals	0.01467	479,782	0.70%
Total	2.10311	479,999	100.00%

Table 3.7b: Mean and SD of Power Conditional on Discrimination for NL3 Change Pattern

	HD	MD	LD
Mean	0.984	0.979	0.829
SD	0.017	0.006	0.025

Table 3.7c: Mean and SD of Power Conditional on Statistic for NL3 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.932	0.936	0.935	0.942	0.883	0.954
SD	0.012	0.011	0.011	0.017	0.045	0.010

Means and standard deviations of power conditional on Discrimination and Statistics conditions are presented in Table 3.7b and 3.7c, respectively. The mean power conditional on discrimination conditions varied in the range of 0.829 to 0.984 and the standard deviation varied in the range of 0.006 to 0.025. The mean power conditional on the statistic for NL3 varied from 0.883 to 0.954 while the standard deviation varied in the range of 0.010 to 0.045. χ^2_{FI} had notably lower power than the other statistics.

Table 3.8a displays results of ANOVA on power for the the NL4 change pattern in which change was Δ = 0.5, 1.0 across the three occasions. The main effect of Discrimination was observed to account for maximum variation with η^2 = 89.16%. Effect sizes of all other main effects, two-way interactions, and error was less than 5%.

Mean and standard deviation of power conditional on the significant main effect of Discrimination under the NL4 condition are depicted in Table 3.8b. The mean varied from 0.727 to 0.991 and the standard deviation varied in the range of 0.007 to 0.034.

Table 3.8a: Results of ANOVA with 2-Way Interactions on Power for the NL4 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.0649	7	1.46%
Discrimination	3.9739	2	**89.16%**
Information/Peakedness	0.0451	1	1.01%
Statistic	0.0898	5	2.02%
$\theta \times$ Discrimination	0.0455	14	1.02%
$\theta \times$ Information	0.0395	7	0.89%
$\theta \times$ Statistic	0.0223	35	0.50%
Discrimination × Information	0.0377	2	0.85%
Discrimination × Statistic	0.046	10	1.03%
Information × Statistic	0.0107	5	0.24%
Residuals	0.0814	479,911	1.83%
Total	4.4568	479,999	100.00%

Table 3.8b: Mean and SD of Power Conditional on Discrimination for NL4 Change Pattern

	HD	MD	LD
Mean	0.991	0.958	0.727
SD	0.007	0.012	0.034

Table 3.9a summarizes ANOVA results on power for the NL5 change pattern in which change was Δ = 0.5, 1.5 across the three occasions. Maximum variation was accounted for by the main effect of Discrimination (η^2 = 45.57%), followed by the type of statistic (η^2 = 11.70%). The two-way $\theta \times$ Statistic interaction accounted for 10.09% of variation and the three-way $\theta \times$ Discrimination × Statistic interaction accounted for

10.01% of the total variation. Other main effects, interactions, and error accounted for less than 5% of variation.

Table 3.9a: Results of ANOVA with 3-Way Interactions on Power for NL5 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.005494	6	1.52%
Discrimination	0.164784	2	**45.57%**
Information/Peakedness	0.007447	1	2.06%
Statistic	0.042315	5	**11.70%**
θ × Discrimination	0.014127	12	3.91%
θ × Information	0.003641	6	1.01%
θ × Statistic	0.036494	30	**10.09%**
Discrimination × Information	0.01517	2	4.20%
Discrimination × Statistic	0.008769	10	2.42%
Information × Statistic	0.001954	5	0.54%
θ × Discrimination × Information	0.005789	12	1.60%
θ × Discrimination × Statistic	0.036193	60	**10.01%**
θ × Information × Statistic	0.008082	30	2.23%
Discrimination × Information × Statistic	0.006593	10	1.82%
Residuals	0.004764	419,808	1.32%
Total	0.361616	419,999	100.00%

Table 3.9b: Mean and SD of Power Conditional on Discrimination for NL5 Change Pattern

	HD	MD	LD
Mean	0.991	0.994	0.938
SD	0.013	0.002	0.011

Table 3.9c: Mean and SD of Power Conditional on Statistic for NL5 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.977	0.979	0.979	0.979	0.946	0.986
SD	0.005	0.005	0.005	0.009	0.032	0.004

The means and standard deviations of significant main effects for NL5 are presented in Table 3.9b, and 3.9c. Mean power for Discrimination varied from 0.938 to 0.991 while the standard deviation varied in the range of 0.002 to 0.013 as depicted in Table 3.9b. Mean power under different statistics condition varied in the range of 0.946 to 0.986 and standard deviation ranged from 0.004 to 0.032, as reflected in Table 3.9c. The means for 2- and 3-way interactions are presented in Appendix Tables A5 and A6, respectively. Figures 3.3 and 3.4 depict the significant 2- and 3-way interactions for NL5.

For the 2-way interaction of $\theta \times$ Statistic, it can be seen that mean power for all statistics was slightly higher at θ levels below 0.0 compared to that at θ levels above 0.0, except for χ^2_{FI}. The 2-way interaction plot shows that mean power remained consistent across θ range, but χ^2_{FI} seemed to underperform at $\theta = -2.0$, compared to other θ levels, resulting in a significant interaction.

Figure 3.4 depicts a significant $\theta \times$ Discrimination × Statistic interaction for NL5. In high and medium discrimination conditions, χ^2_{FI} resulted in lower power compared to other statistics, with other statistics resulting in very similar power. In low discrimination condition, however, the differences in power of other statistics increased slightly.

Results of ANOVA on power for the last change pattern, NL6, in which change was introduced in the magnitude of $\Delta = 1.0, 1.5$ across the three occasions are presented in Table 3.10a. The maximum variation was accounted for by the main effect of Statistic

Figure 3.3: **2-Way θ × Statistic Interaction for NL5 Change Pattern**

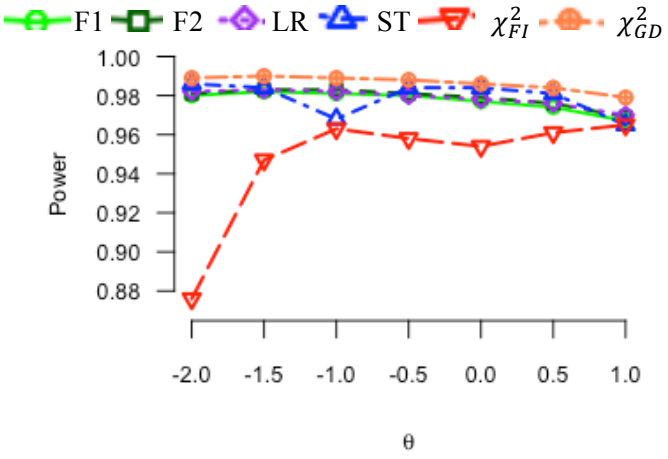
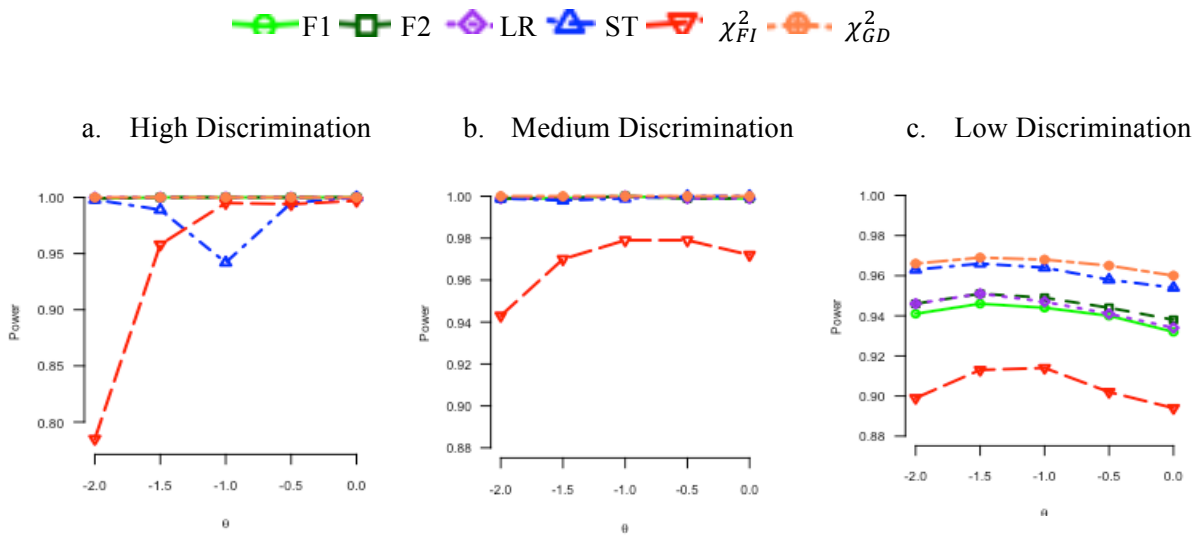


Figure 3.4: **3-Way θ × Discrimination × Statistic Interaction for NL5 Change Pattern**



(η^2 = 33.93%) followed by that of Discrimination (η^2 = 19.24%). The two-way θ ×Statistic interaction accounted for 11.80% of variation and the Information × Statistic interaction accounted for 7.16% variation. The three-way θ × Discrimination × Statistic interaction was observed to account for 5.05% variation and the Discrimination × Information × Statistic interaction accounted for 6.9% variation. Variability due to all other main effects, interactions, and error remained under 5%.

Table 3.10a: Results of ANOVA with 3-Way Interaction on Power for NL6 Change Pattern

Source of Variation	Sum of Squares	Degrees of Freedom	η^2
θ	0.0004485	5	1.60%
Discrimination	0.0054077	2	**19.24%**
Information/Peakedness	0.0007205	1	2.56%
Statistic	0.0095375	5	**33.93%**
θ × Discrimination	0.0006072	10	2.16%
θ × Information	0.0003153	5	1.12%
θ × Statistic	0.0033179	25	**11.80%**
Discrimination × Information	0.0020137	2	**7.16%**
Discrimination × Statistic	0.0004276	10	1.52%
Information × Statistic	0.0003789	5	1.35%
θ × Discrimination × Information	0.0001657	10	0.59%
θ × Discrimination × Statistic	0.0014192	50	**5.05%**
θ × Information × Statistic	0.0010635	25	3.78%
Discrimination × Information × Statistic	0.001938	10	**6.90%**
Residuals	0.000345	359,834	1.23%
Total	0.0281062	359,999	100.00%

Table 3.10b: Mean and SD of Power Conditional on Discrimination for NL6 Change Pattern

	HD	MD	LD
Mean	0.997	0.997	0.987
SD	0.003	0.003	0.001

Table 3.10c: Mean and SD of Power Conditional on Statistic for NL6 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Mean	0.996	0.996	0.996	0.997	0.979	0.998
SD	0.001	0.001	0.001	0.001	0.011	0.0005

Table 3.10b and 3.10c present means and standard deviations of the significant main effects. Table 3.10b shows that mean power varied from 0.987 to 0.997 under the discrimination conditions while the standard deviations varied in the range of 0.001 to 0.003. Mean power for 2- and 3-way interactions can be found in Appendix Tables A7, A8, A9 and A10.The significant 2- and 3-way interactions for NL6 are depicted in Figures 3.5, 3.6, 3.7 and 3.8, respectively.

From Figure 3.5 depicting the 2-way $\theta \times$ Statistic interaction, it can be seen that mean power for all statistics remained consistently high across all θ levels for except for χ^2_{FI} which resulted in lower power at θ levels below 0.0. Mean power for the statistics ranged from 0.979 to 0.998. Standard deviations for mean power for different statistics under various θ levels ranged from 0.003 to 0.025 (Appendix Table A7).

Figure 3.5: 2-Way $\theta \times$ Statistic Interaction for the NL6 Change Pattern

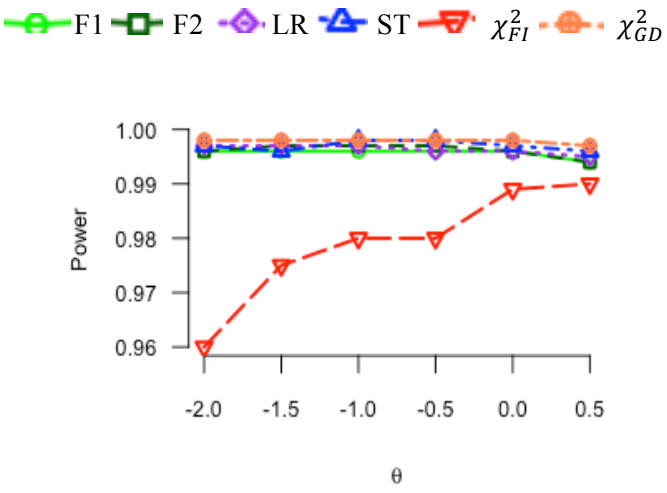


Figure 3.6 presents the 2-way Discrimination × Information interaction. For high and medium discrimination, both flat and peaked item banks resulted in similar power. However, in the low discrimination condition, the peaked bank resulted in higher power for NL6. Mean power was observed to be very close for high and medium discrimination

conditions with that being 0.997 and 0.998, respectively for flat tests and 0.997 for peaked tests. In case of the low discrimination condition, peaked tests resulted in mean power of 0.993 and flat tests resulted in mean power of 0.980 (Appendix Table A8).

Figure 3.6: 2-Way Discrimination × Information Interaction for the NL6 Change Pattern

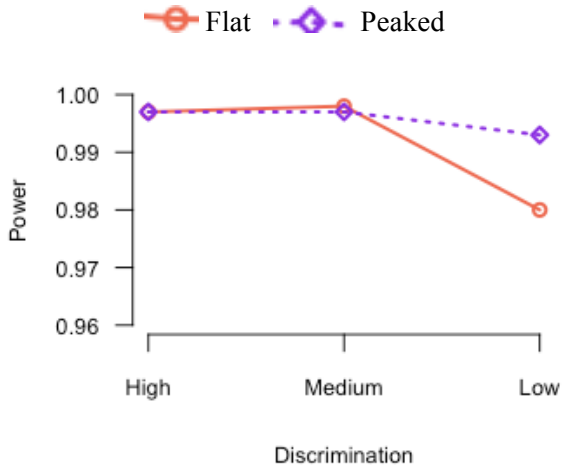


Figure 3.7 presents the 3-way θ × Discrimination × Information interaction for NL6. From this interaction plot, it can be seen that χ^2_{FI} underperformed across all discrimination conditions. In high and medium discrimination conditions, however, χ^2_{FI} resulted in higher power as θ increased. In contrast, in the low discrimination condition, χ^2_{FI} seemed to underperform across the θ range. The curves representing different statistics almost overlapped for high and medium discrimination, whereas they were slightly apart in the low discrimination condition.

Figure 3.8 presents 3 way θ × Discrimination × Information interaction, found significant for NL6. In the high and medium discrimination conditions, flat item banks resulted in higher power compared to the peaked item banks. The item banks also resulted in similar power across all the statistics, with the exception of χ^2_{FI}. In the low discrimination

condition, this trend was reversed with peaked banks resulting in higher power compared to the flat banks, resulting in a significant interaction effect.

Figure 3.7: **3-Way θ × Discrimination × Statistic Interaction for the NL6 Change Pattern**

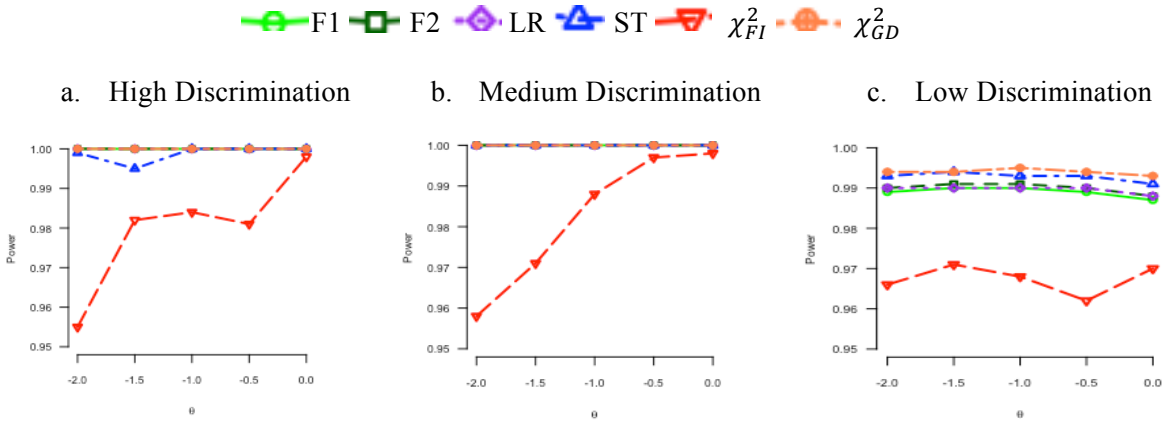
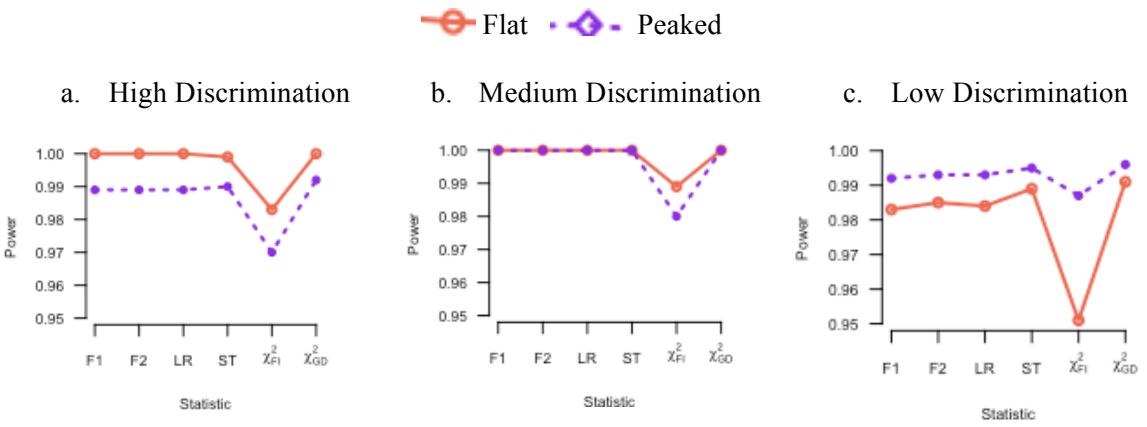


Figure 3.8: 3-Way Discrimination × Information × Statistic Interaction for NL6 Change Pattern



Overall, ANOVA results indicated that the factors that consistently influenced the variation in Type I error and power were discrimination, type of statistic, and θ in some cases. The two-way θ × Statistic and the three-way θ × Discrimination × Statistic interaction was also observed to be influencing significant proportions of variation for

change patterns of L3, NL5 and NL6, i.e., the change conditions consisting of high amount of change.

Effect of θ

Figure 3.9 depicts mean Type I error and power for all change patterns at different θ levels. In Figure 3.9a, the dark black line represents observed mean Type I error versus the dashed line at 0.05 for comparison. Figure 3.9b represents observed mean power conditional on θ. In Figure 3.9c observed power is depicted for linear patterns of change and in Figure 3.9d, observed power is depicted for non-linear patterns of change.

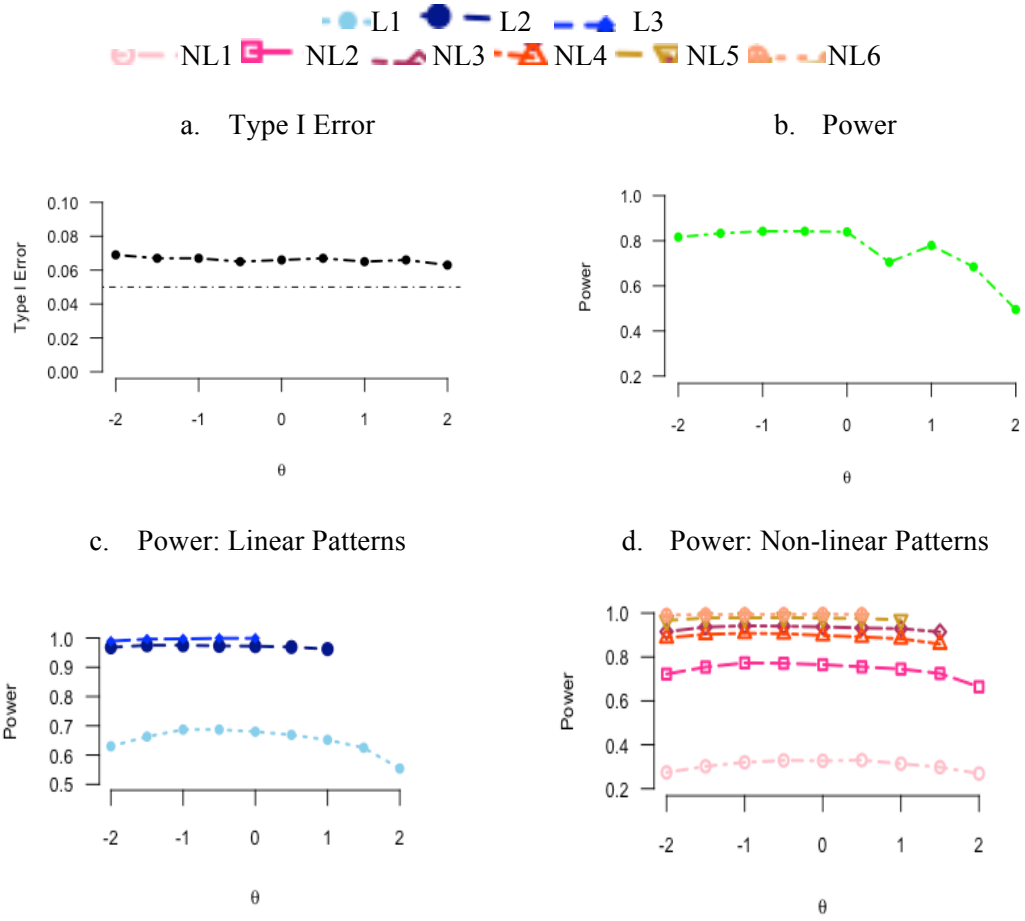
Figure 3.9a shows that observed Type I error was around 0.7 across all θ levels. Observed power as depicted in Figure 3.9b was around 0.8 across θ levels –2.0 to 0. Power dropped at higher θ levels (0.5 to 2.0). Among the linear change patterns (L1, L2 & L3), observed power remained around 0.95 to 1.0 for L2 and L3 whereas it ranged from 0.6 to 0.7 for L1.

Among the non-linear change patterns (NL1 through NL6), lowest power was observed in the range of 0.2 to 0.3 under the NL1 condition. It remained around 0.7 for NL2 and ranged from 0.88 to 1.0 for the remaining patterns of change. Low amount of change in NL1 (Δ = 0, 0.5) resulted in low power compared to other change patterns.

In Figure 3.9c, the curves for observed power end at θ = 1 for L2 and θ = 0 for L3. Similarly, in Figure 3.9d, the curves end at θ = 1.5 for NL3 and NL4, at θ = 1 for NL5, and at θ = 0.5 for NL6. This is due to the fact that these change patterns were not introduced at those missing θ levels in order to control the upper limit at θ = 3. This is also a reason why the observed power dropped off at θ = 0.5 and beyond in Figure 3.9b. All change patterns, especially the patterns representing high level of change, were not introduced

above $\theta = 0$, which resulted in lower means at $\theta = 0.5$ and above compared to means at and below $\theta = 0$.

Figure 3.9: Mean Type I Error and Power Conditional on θ



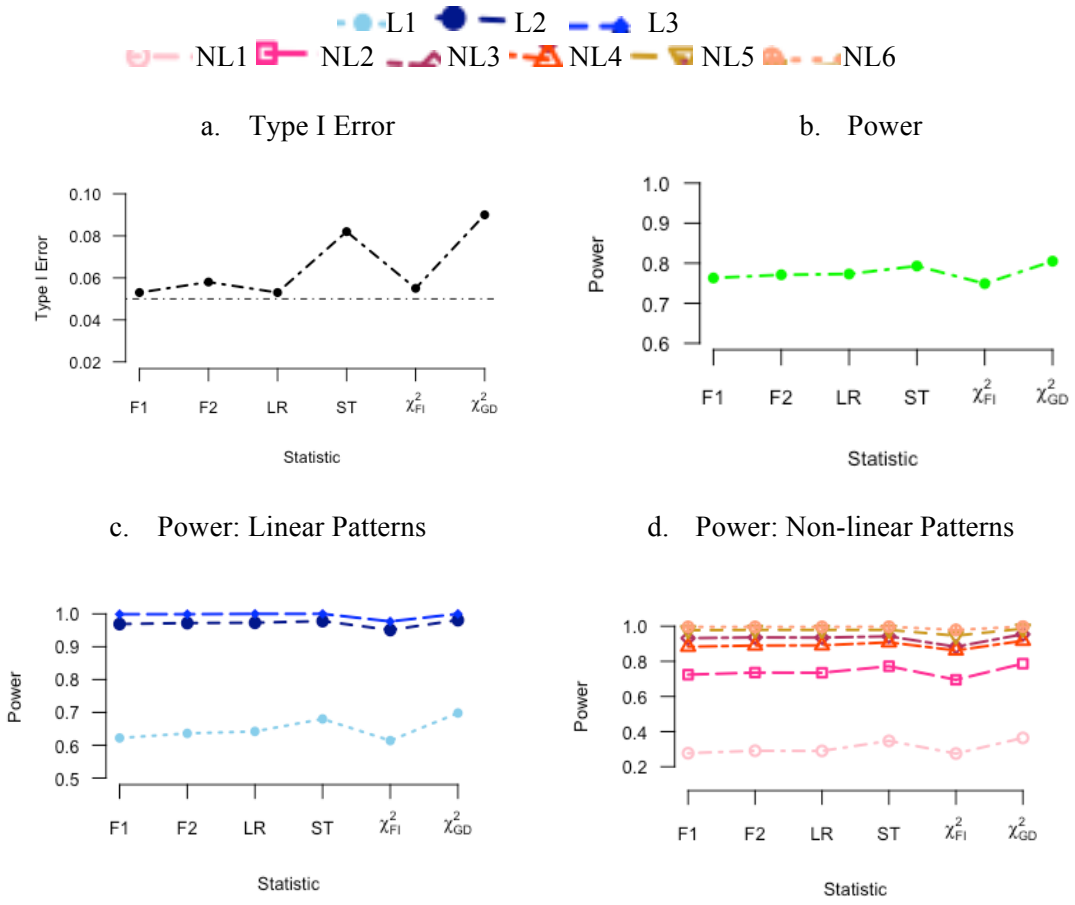
For most change patterns, observed η^2 remained below 2%, except for the L3 condition, in which η^2 was observed to be 5.18% (Table 3.4a). Overall, Type I error was slightly higher at the negative end of θ. Power was observed to drop off slightly at lower and upper θ extremes and remained relatively consistent for middle θ values.

Effect of Statistic

Figure 3.10 shows the effect of statistic on Type I Error and Power under various change conditions. 3.10a depicts observed Type I Error for all statistics against desired Type I Error of 0.05. Marginal Type I error (Figure 3.10a) remained around 0.05 for F1,

F2, LR, and χ^2_{FI} statistics. As indicated by Figure 3.10a and presented in Table 3.1b, mean Type I error was observed to be about .08 for ST and about .09 for χ^2_{GD} statistic. For Type I Error, η^2 accounted for by type of Statistic was 80.93% (Table 3.1a).

Figure 3.10: Mean Type I Error and Power Conditional on Statistic



Observed mean power across all change patterns (Figure 3.10a) varied from 0.7 to 0.8 for all statistics. Observed power was slightly higher for ST and χ^2_{GD} statistics compared to the F1, F2, LR, and χ^2_{FI} statistics, although they both had higher Type I errors.

Among the linear change patterns (Figure 3.10c), power ranged from 0.6 to 0.7 for L1 and from 0.96 to 1.0 for L2 and L3. η^2 for Statistic in the L1 change pattern condition was 1.90% (Table 3.2a). η^2 for Statistic under the L2 and L3 change patterns was 5.72%

(Table 3.3a) and 28.28% (Table 3.4a), respectively. For the linear patterns, χ^2_{FI} was observed to display slightly lower power compared to the other statistics.

Observed power under the non-linear patterns of change (Figure 3.9d) ranged from 0.2 to 0.4 for NL1, remained around 0.7 for NL2 and ranged from 0.84 to 1.0 for NL3, NL4, NL5, and NL6. In terms of power of the statistics, the trend for non-linear change patterns remained similar to that of linear change patterns. Observed power remained consistent across all the types of statistics under non-linear change pattern conditions. Observed power for the χ^2_{FI} statistic was slightly less than that for other statistics. The type of statistic was observed to be contributing significantly toward total variation in NL1, NL3, NL5, and NL6 conditions in which η^2 was 6.44%, 6.90%, 11.70% and 33.93%, respectively (Table 3.5a, 3.7a, 3.9a and 3.10a). η^2 was observed to be around 2% for NL2 and NL4 conditions (Table 3.6a and 3.8a). Under the NL2 and NL4 conditions, almost all the variation in observed power was accounted by Discrimination (91.02% and 89.16%, respectively), resulting in unsubstantial proportions of variance attributable to other effects in the model. The small remainder proportion after accounting for Discrimination was divided among other factors and interactions, deeming all other factors resulting in very small contribution to total variation.

Figure 3.11 displays mean Type I error and power conditional on the type of statistic and θ. In Figure 3.11a, it can be seen that Type I error for χ^2_{FI} and F1 statistics remained very close to 0.05, except for θ = 1.5 and θ = 2.0 at which Type I error for χ^2_{FI} was about 0.6 and 0.7, respectively. For F2 and LR statistics, the Type I error ranged from 0.055 to 0.06. Both ST and χ^2_{GD} displayed higher Type I error. That of ST ranged from 0.07 to 0.09 and that of χ^2_{GD} ranged from 0.08 to 0.09. Figure 3.11b shows that observed power

decreased from θ = 1.0 as θ increased. Mean power for the statistics ranged from 0.75 to 0.85 from θ = –2 to θ = 1.0. Observed power was 0.65 at θ = 1.5 and around 0.5 at θ = 2. This effect was observed, as noted previously, because high change patterns were not introduced at high θ levels in order to limit θ at 3.0 after change. It is interesting to note that although ST and χ^2_{GD} had higher Type I error than the other statistics, their power did not differ much.

Figure 3.11: Mean Type I Error and Power Conditional on Statistic and θ

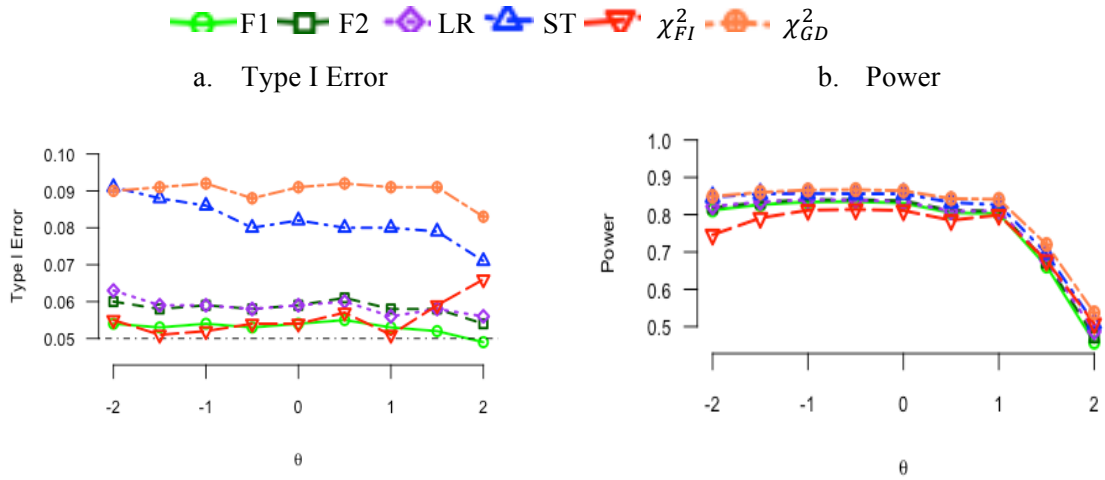


Figure 3.12 depicts mean power conditional on the type of statistic and θ for all change patterns. For most of the statistics, power dropped slightly at the lower and upper ends of θ. For χ^2_{FI}, however, observed power was relatively lower at θ = –2 compared to other statistics, but power increased as θ increased for some change patterns. The relative performance of the statistics in detecting power remained consistent across all change patterns. χ^2_{GD} displayed highest power across all conditions followed by ST. After ST, F2 and LR showed consistent performance across all change patterns. Lastly, χ^2_{FI} and F1 showed slightly lower power across all conditions. However, differences in the power of these statistics were negligible (except forχ^2_{FI}), especially under high change conditions

compared to low change conditions. For example, the maximum difference between observed power of χ^2_{FI} and χ^2_{GD} was about 0.2 under low and medium change conditions (L1, NL1, NL2). However, the differences reduced for high change conditions (L2, L3, NL2, NL3, NL4, NL5, NL6).

Another noticeable trend in Figure 3.12 is differences in observed power for linear vs. non-linear patterns. For the same total magnitude of change (e.g. L1 vs. NL2 or L2 vs. NL5), observed power was higher under non-linear change patterns than that in linear change patterns. These, too, were negligible for high levels of change conditions.

Effect of Discrimination

Figure 3.13 displays mean Type I error and power conditional on discrimination for all change patterns. It can be seen from Figure 3.13a that Type I error remained mostly consistent across the three discrimination conditions, with that for the low discrimination condition being slightly lower than Type I error for high and medium discrimination. Observed η^2 for Discrimination for Type I error was 0.48%.

As expected, mean power was the highest in the high discrimination condition (Figure 3.13b), followed by that in the medium discrimination condition and lastly in the low discrimination condition. Observed power was about 0.9 for high discrimination, about 0.8 for medium discrimination and about 0.6 for the low discrimination condition.

Figure 3.13c shows that among the linear change patterns, observed power for L1 was about 0.9 and that of L2 and L3 was about 1.0 for the high discrimination condition. Power declined significantly for the L1 change pattern under medium and low discrimination conditions. For L2 and L3 change patterns, observed power remained around 1.0 under medium discrimination, with that of L2 decreasing slightly under the low

Figure 3.12: **Mean Power Conditional on Statistic and θ for Different Patterns of Change**

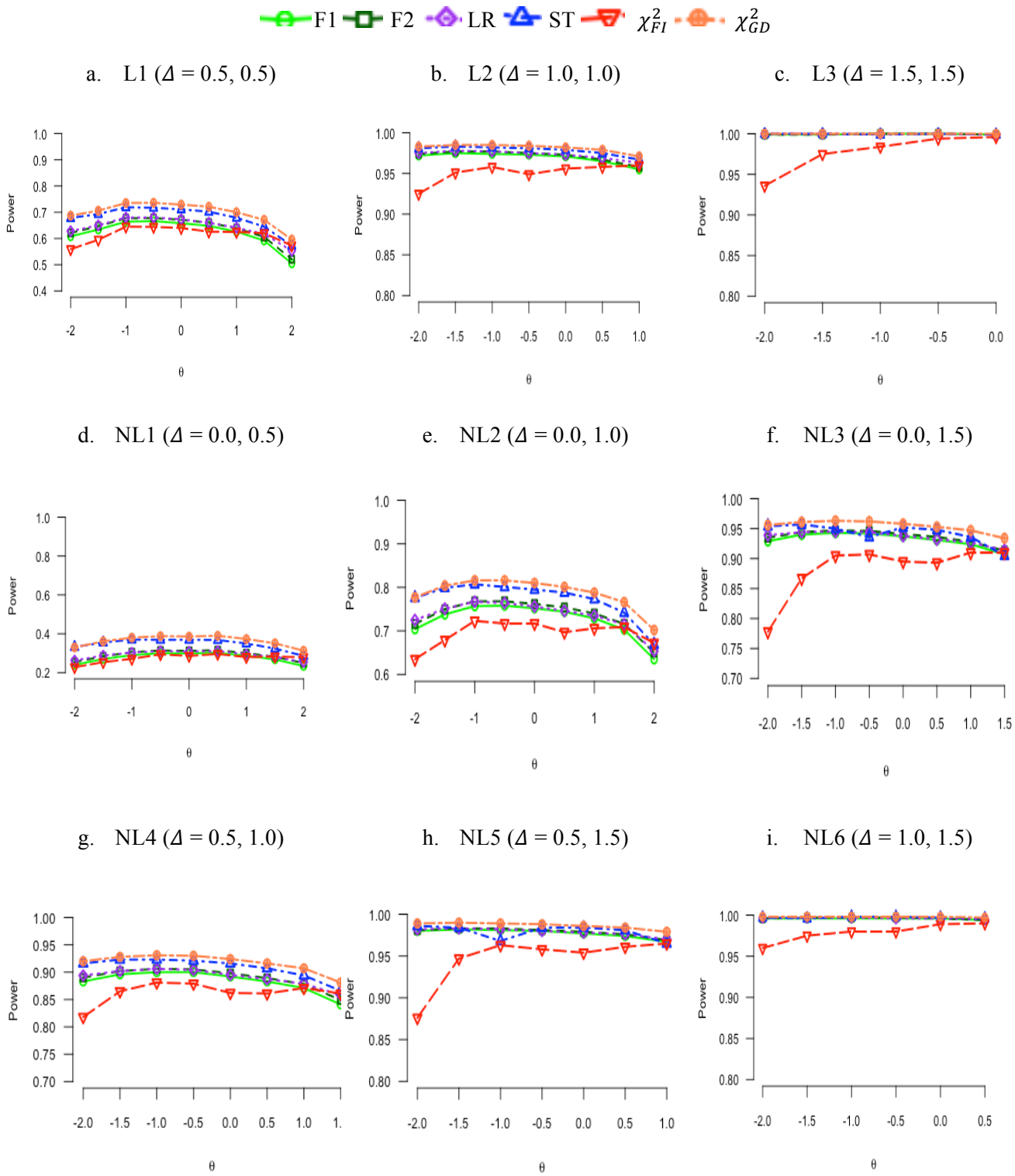
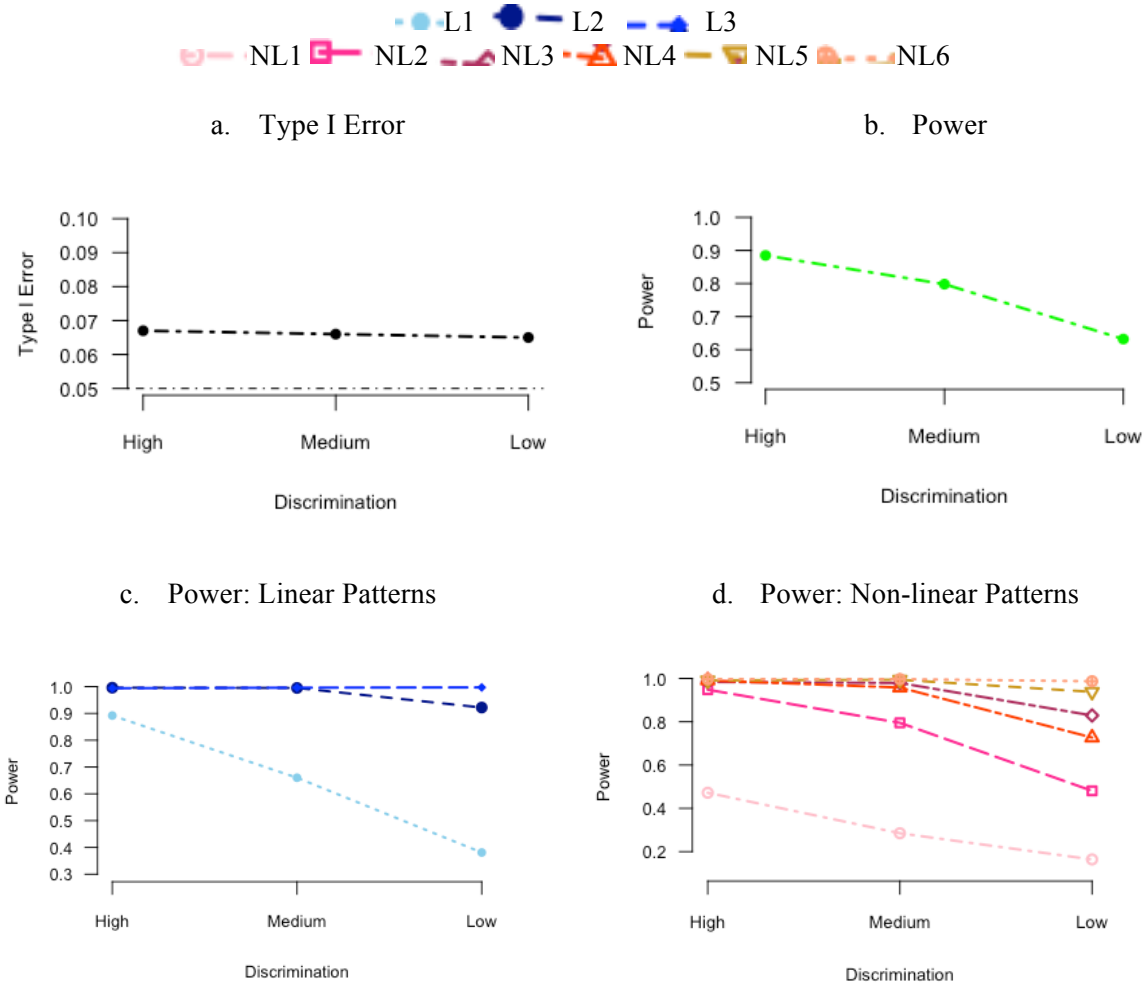


Figure 3.13: Mean Type I Error and Power Conditional on Discrimination



discrimination condition. Discrimination was found to account for a significant proportion of variability in the ANOVA framework. Observed η^2 for L1, L3, and L3 was 90.83%, 74.12%, and 0.74%, respectively (Table 3.2a, 3.3a and 3.4a).

Figure 3.13d shows a similar trend for non-linear change patterns. Observed power was about 1.0 for all non-linear change patterns in the condition of high discrimination, with an exception of the NL1 pattern in which the power was about 0.48. Power dropped further to about 0.2 for NL1 under medium and low discrimination. Observed power for NL3, NL4, NL5, and NL6 remained at 1.0 under the medium discrimination condition with that for NL2 being about 0.8. Power for NL2 dropped to about 0.5 in the low discrimination

condition. Power dropped slightly for the rest of the change patterns, as well, in the low discrimination condition and remained in the range of 0.75 to 1.0. Differences in observed power due to discrimination reduced with high levels of change. For non-linear change patterns, discrimination accounted for a significant proportion of total variability in the observed power, with η^2 = 84.91%, 91.02%, 70.58%, 89.16%, 45.57%, and 19.24% for NL1 through NL6, respectively (Table 3.5a, 3.6a, 3.7a, 3.8a, 3.9a and 3.10a).

Figure 3.14 displays mean Type I error and power conditional on discrimination and θ. From Figure 3.14a, it can be seen that Type I error remained in the range of 0.06 to 0.07 across all θ levels. The differences in Type I error for discrimination conditions were negligible. η^2 for the θ × Discrimination interaction remained below 5% for Type I error and Power under all the change patterns.

Figure 3.14: Mean Type I Error and Power Conditional on Discrimination and θ

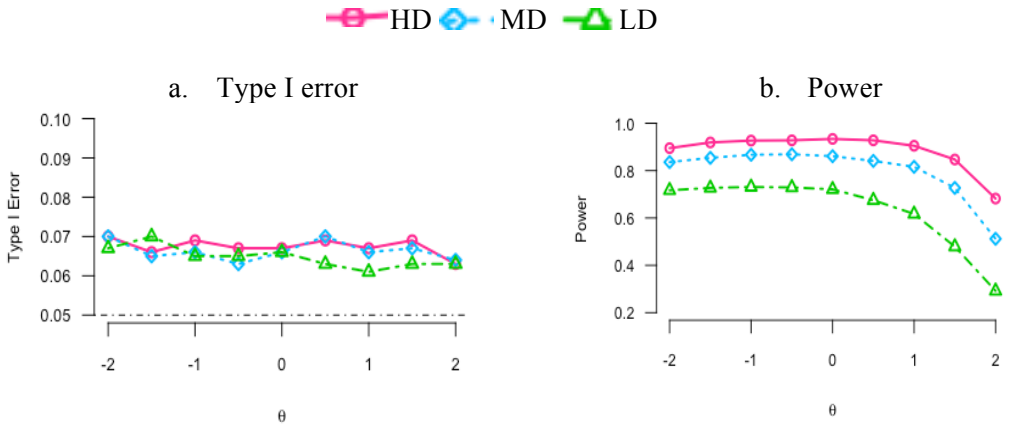


Figure 3.14b shows that the high discrimination condition resulted in maximum observed power across all θ conditions ranging from 0.92 to 0.7. Observed power for medium discrimination was also close to that under high discrimination, ranging from 0.84 to 0.55.

Observed power for low discrimination ranged from 0.7 to 0.23 across the θ levels.Figure 3.15 displays mean power conditional on discrimination and θ for different patterns of change. The $\theta \times$ Discrimination interaction did not contribute substantially toward the total variation in ANOVA, and the observed η^2 remained below 5% for all change patterns. Curves representing observed power under the conditions of high and medium discrimination were very close and almost overlapped when the amount of change was high, particularly for change patterns L2, L3, NL3, NL4, NL5, and NL6.

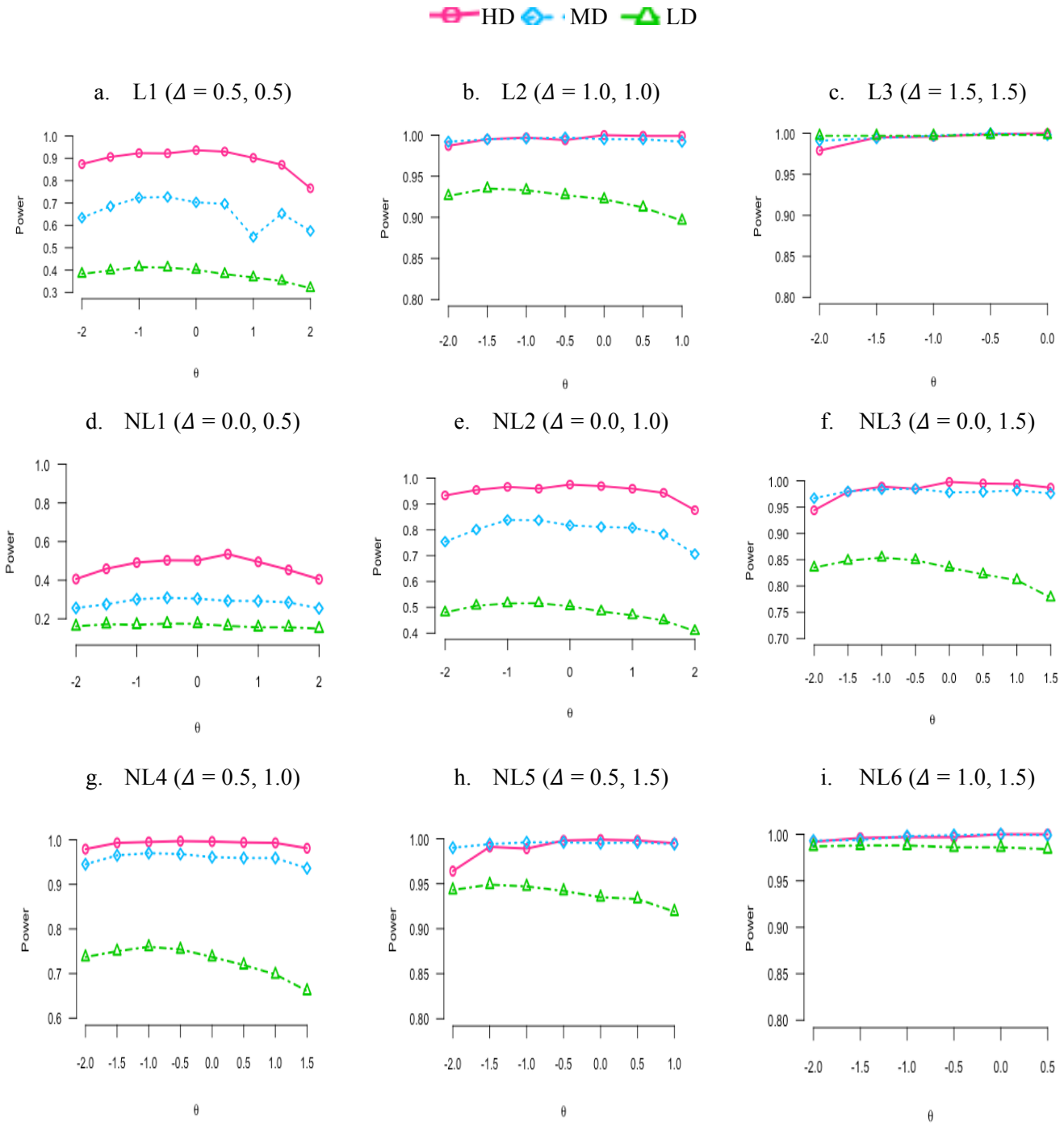
At very high levels of change, namely L3 (Figure 3.15c) and NL6 (Figure 3.15i), the curve representing low discrimination also approximated that of high and medium discrimination. As displayed in Figure 3.15, even though the amount of linear or non-linear change increased, the low discrimination condition resulted in reasonable power of approximately 0.7 (e.g., Figure 3.15g).

At small levels of change (NL1; Figure 3.15d), high discrimination resulted in highest observed power (around 0.4 to 0.5) followed by medium (around 0.22 to 0.3) and low discrimination (about 0.15). Overall, differences in observed power due to discrimination diminished as amount of linear or non-linear change increased.

Effect of Information

Figure 3.16 shows the effect of bank information/peakedness on mean Type I error and power. It can be seen in Figure 3.16a that Type I error was slightly less for peaked banks compared to flat banks. Conversely, power was slightly higher for peaked

Figure 3.15: Mean Power Conditional on Discrimination and θ for Different Patterns of Change



banks compared to flat banks. Among the linear, as well as the non-linear patterns, power was slightly higher for peaked compared to flat banks but this difference diminished as the

amount of change became higher. None of the differences resulted in important eta-squared values.

Figure 3.17 displays mean Type I error and power for peaked and flat banks conditional on θ. Flat banks resulted in higher Type I error compared to peaked banks across all levels of θ, except at $\theta = -1$ and $-0.5$where the two curves overlapped. Peaked banks also resulted in marginally higher power across all levels of θ, except the extremes at $\theta = -2$ and at $\theta = 2$ where flat banks resulted in slightly more power. The observed η^2 for the $\theta \times$ Information interaction was below 5% for Type I error as well as for Power under all change pattern conditions.

Figure 3.18 shows mean power conditional on information and θ for different patterns of change. All the change patterns displayed in Figure 3.18 show a consistent trend that CATs from peaked banks resulted in more power than those from flat banks except at the lower and upper end of θ, where power in the flat condition either exceeded power in the peaked condition or remained the same as in the peaked condition. There was no difference in power for CATs from banks with large change (Figure 3.18c and 3.18i), whether change was linear or non-linear.

All the change patterns displayed in Figure 3.18 show a consistent trend that CATs from peaked banks resulted in more power than those from flat banks except at the lower and upper end of θ, where power in the flat condition either exceeded power in peaked condition or remained the same as in peaked condition. There was no difference in power for CATs from banks with large change (Figure 3.18c and 3.18i), whether change was linear or non-linear.

Figure 3.16: Effect of Information on Type I Error and Power

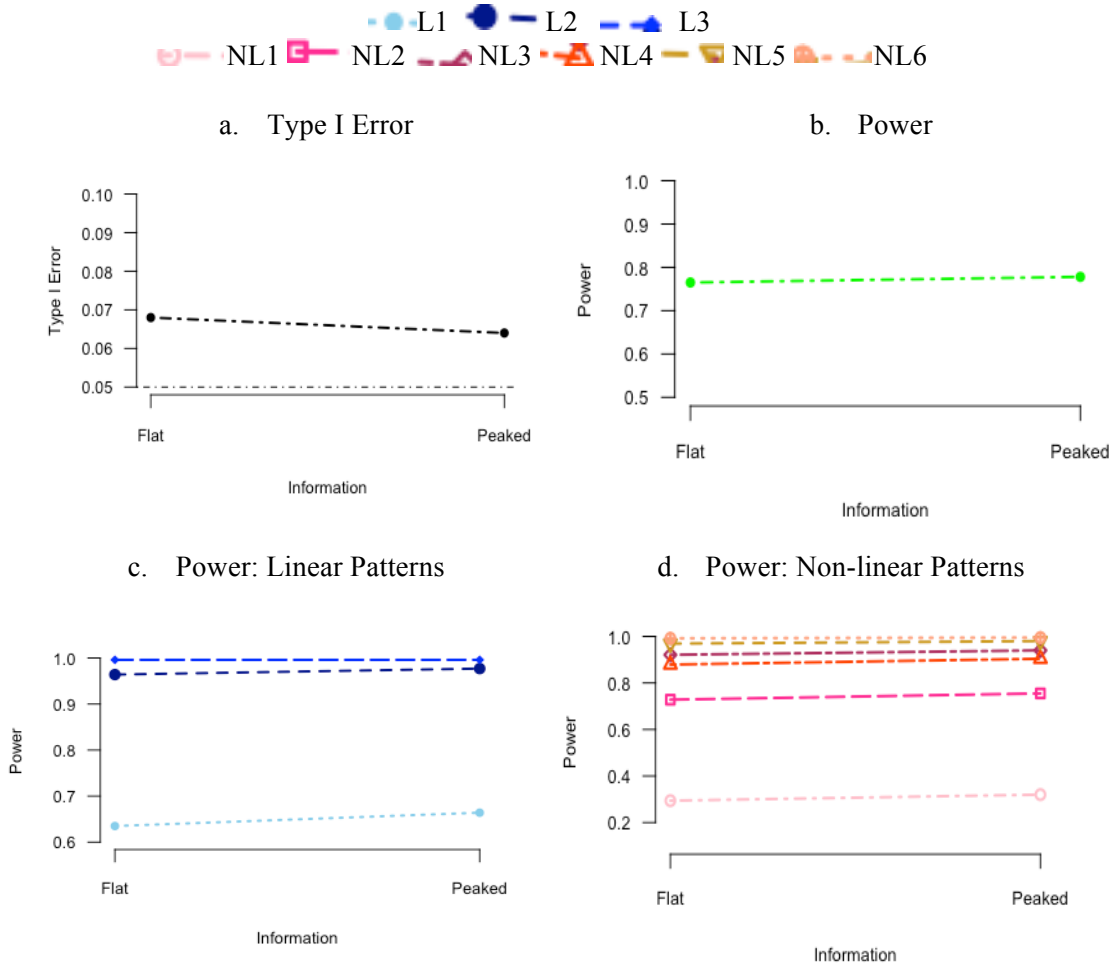


Figure 3.17: Mean Type I Error and Power Conditional on Information and θ

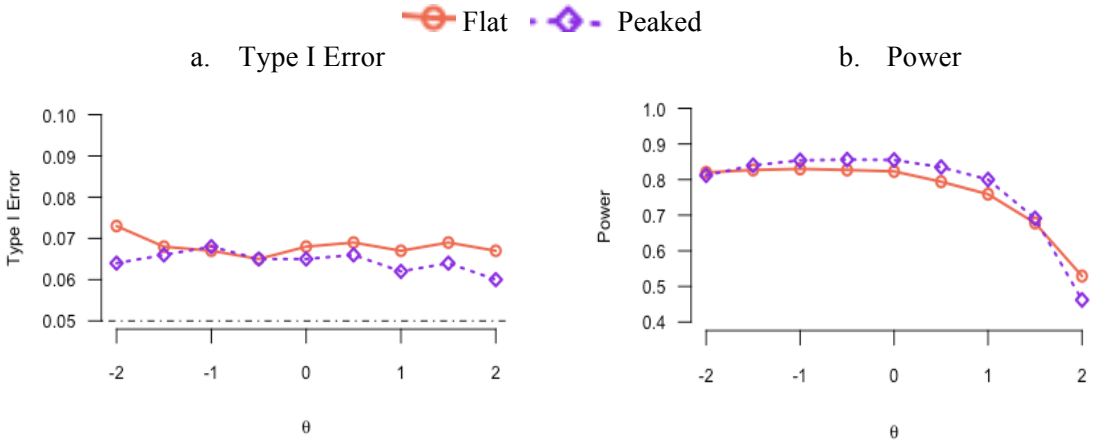
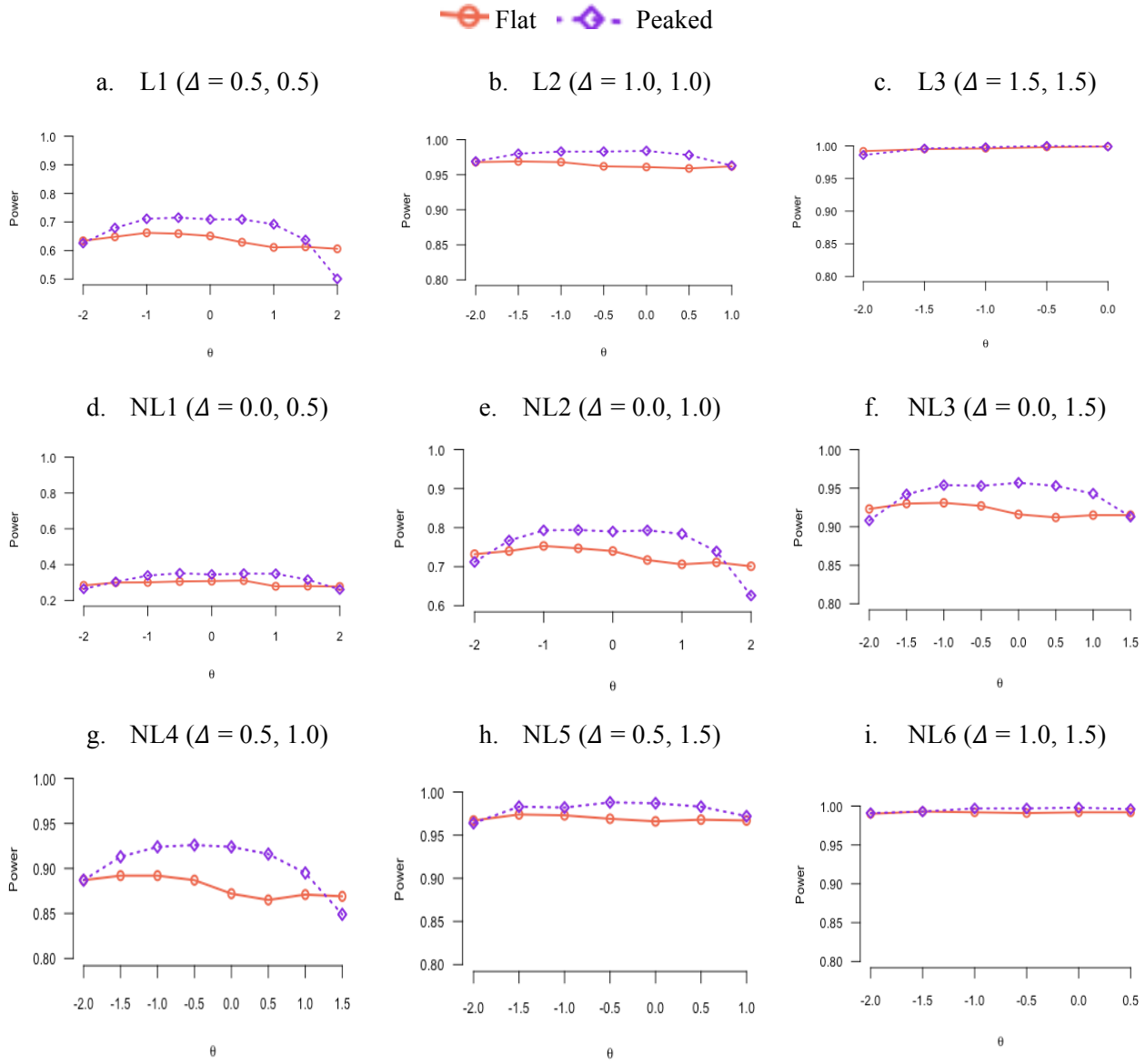


Figure 3.18: Mean Power Conditional on Information and θ for Different Patterns of Change



Effect of Bank Type

Figures 3.19a and 3.19b show the effect of bank type on overall mean Type I error and power. Figure 3.19a shows that Type I error remained consistent, and slightly high, across all six types of item banks. Figure 3.19b shows power decreasing from high discrimination to medium discrimination banks and dropping further for the low discrimination banks. Figure 3.19c shows that power decreased for the L1 change pattern

as the discrimination in item banks decreased. However, power remained consistent across all item banks for L2 and L3 change patterns with high amounts of change. As with the linear patterns, a similar pattern was observed for non-linear patterns in Figure 3.19d. As the amount of change increased, the difference between power due to bank type decreased.

Figure 3.19: Effect of Bank Type on Type I Error and Power

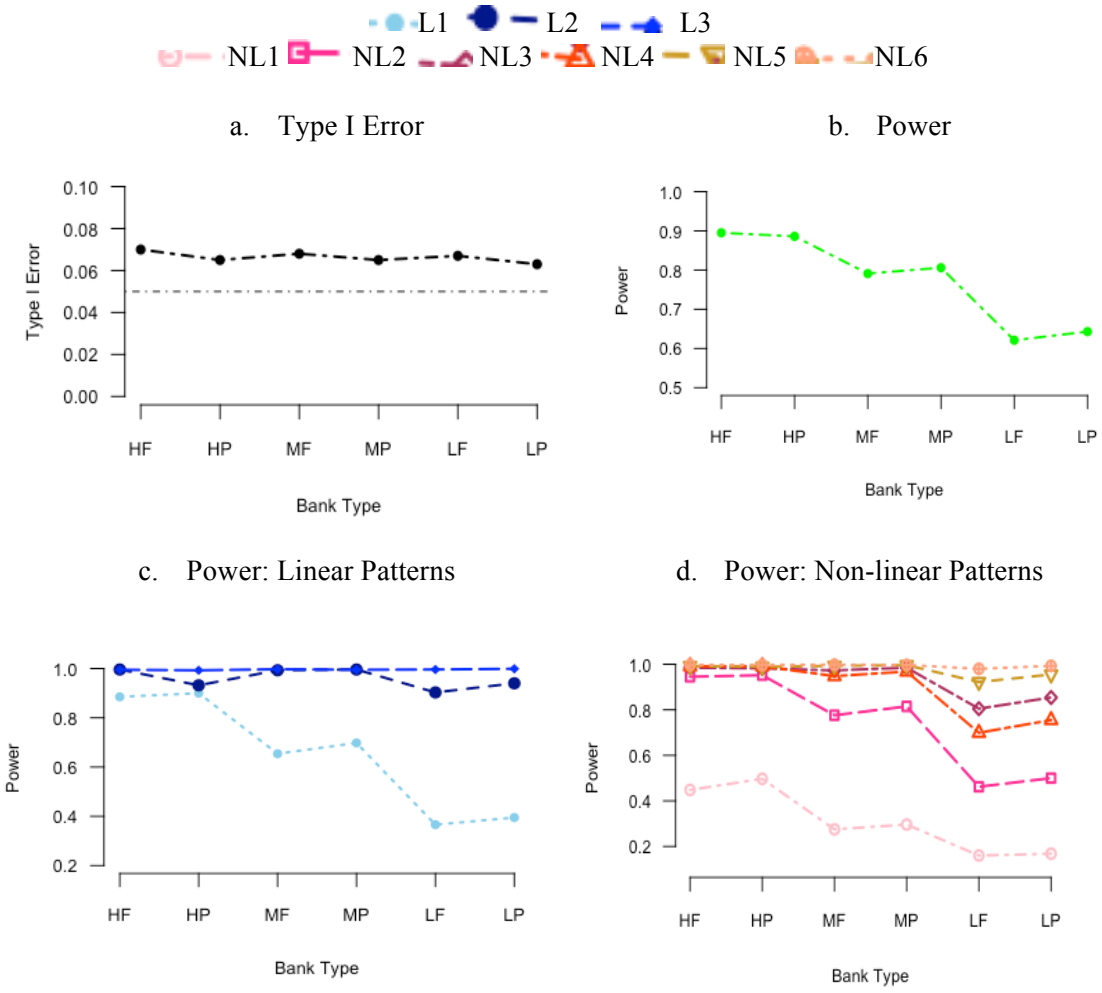
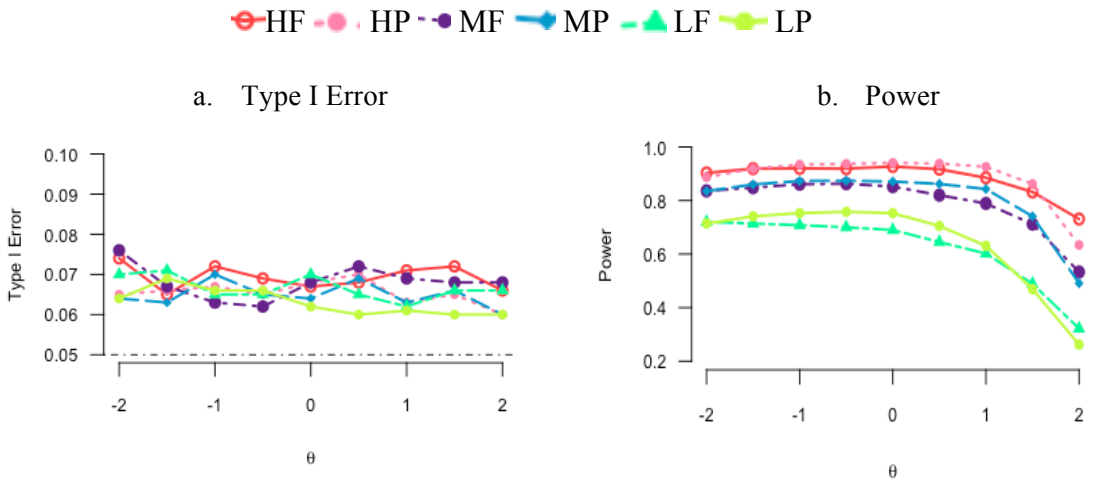


Figure 3.20 shows mean Type I error and power for different bank types conditional on θ. Figure 3.20a shows that Type I error was higher for flat tests compared to peaked

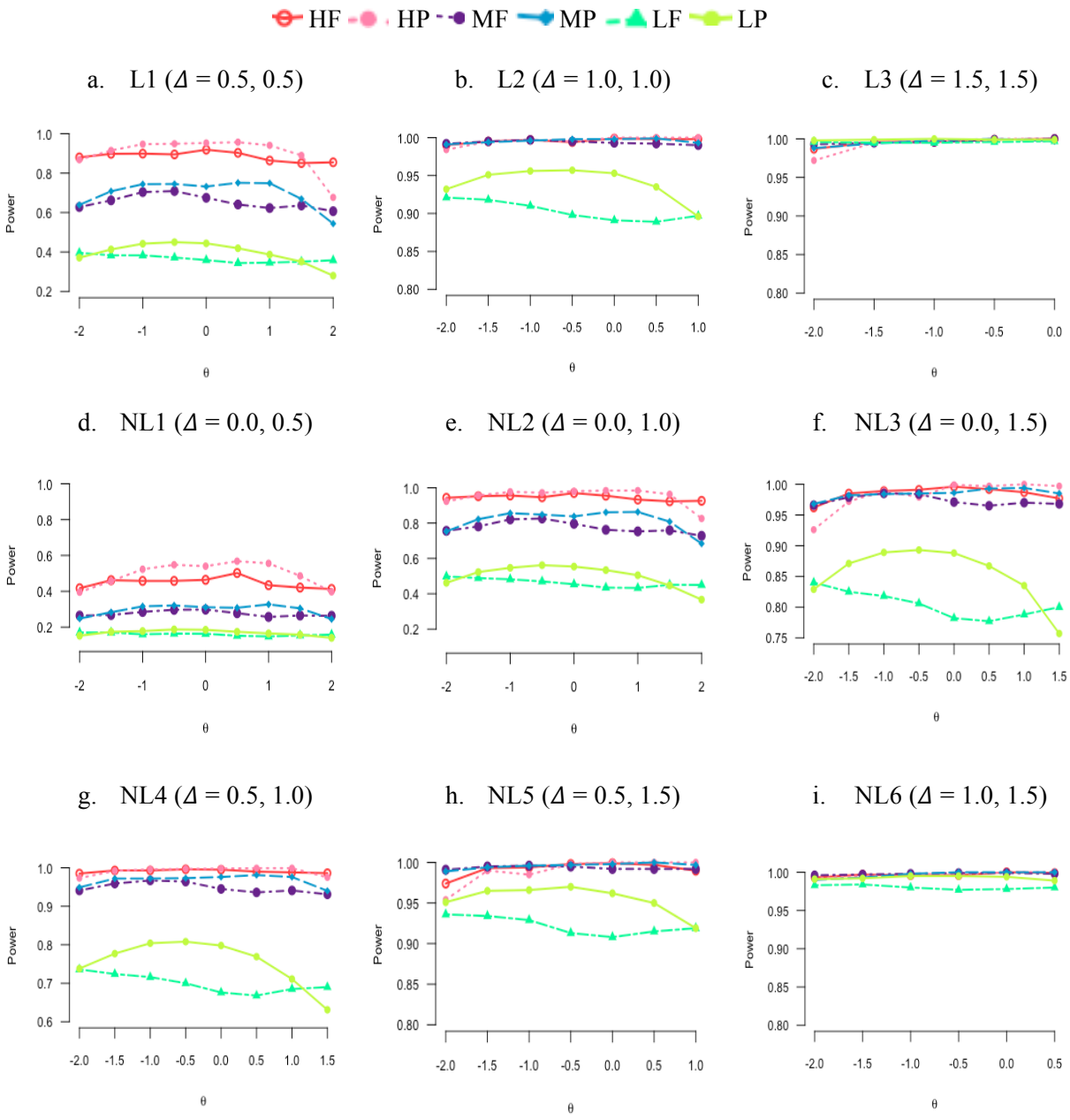
Figure 3.20: Mean Type I Error and Power Conditional on Bank Type and θ



tests at $\theta = -2$, but about the same at $\theta = 2$. All curves representing the different bank types remained essentially consistent across θ levels. In Figure 3.20b, it can be seen that power was the same for flat and peaked tests at $\theta = -2$. However, flat banks had slightly higher power than peaked banks at $\theta = 2$. Peaked banks displayed marginally higher power than flat tests from $\theta = -1.5$ to $\theta = 1.5$. η^2 for all the Discrimination × Information interactions and the θ × Discrimination × Information interactions were below 5%.

Figure 3.21 shows mean power conditional on θ for different bank types and different change patterns. Across all change patterns, peaked item banks resulted in higher power than flat item banks from $\theta = -1.5$ to $\theta = 1.5$. At $\theta = -2$, the two item banks showed similar power and at $\theta = 2$, flat banks resulted in slightly higher power than peaked banks. The differences in power resulting from flat and peaked banks diminished as magnitude of change increased (e.g., Figures 3.21c and 3.21i).

Figure 3.21: Mean Power Conditional on Bank Type and θ for Different Patterns of Change

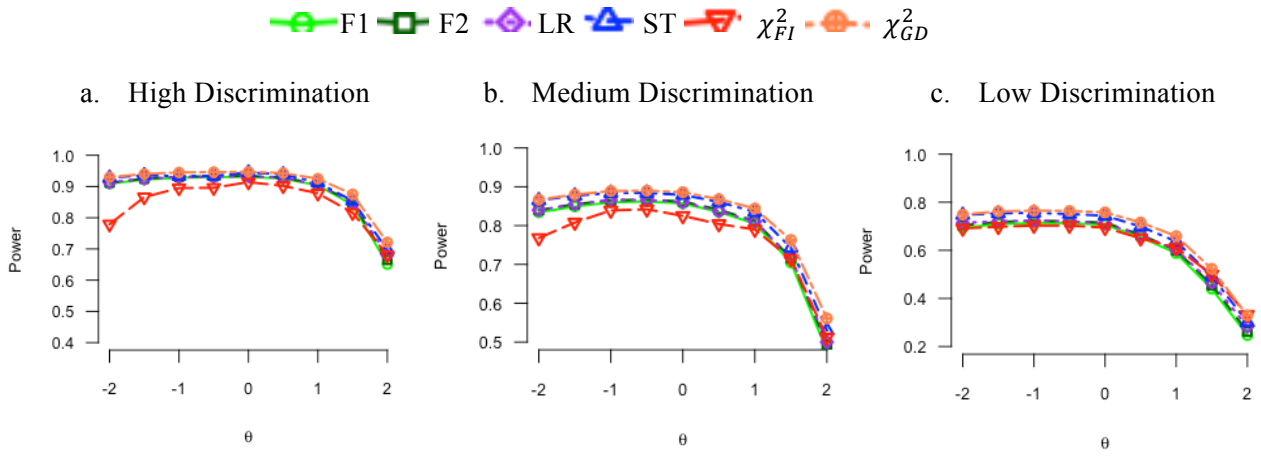


Interaction between θ – Discrimination – Statistic

Figure 3.22 displays the three-way effect of θ × Discrimination × Statistic. This particular interaction was observed to have an effect size of more than 5% for change patterns L3, NL5, and NL6 (Table 3.4a, 3.9a and 3.10a). It can be seen that observed power

was higher in the high discrimination condition followed by medium and low discrimination conditions.

Figure 3.22: Mean Power Conditional on Statistic and θ for Different Discrimination Conditions



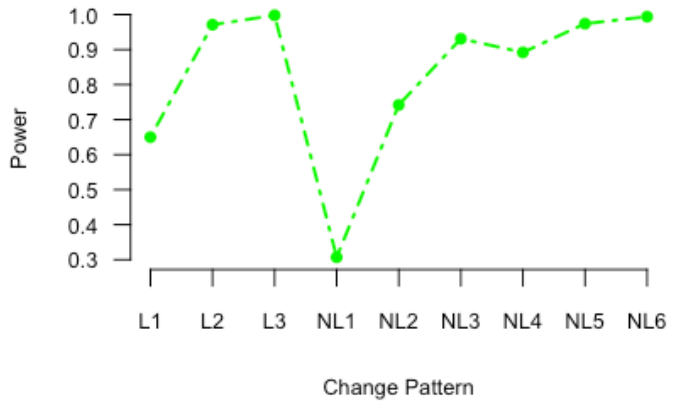
In all the discrimination conditions, χ^2_{FI} underperformed compared to other statistics at θ = –2, but the curve representing χ^2_{FI} crossed those representing other statistics at θ = 2. Although the trend is not apparent when averaged across all change conditions in Figure 3.22, χ^2_{FI} seemed to underperform severely for L3, NL5, and NL6 change patterns under high and medium discrimination conditions in which the 3-way interaction contributed more than 5% variation in the ANOVA.

Effect of Change Patterns

Figure 3.23 displays mean power as a function of change pattern. As can be expected, power is seen to be directly related to change pattern. Larger amounts of change resulted in higher power. For the same amount of change, for example L1 vs.NL2 (total Δ = 1.0) and L2 vs. NL5 (total Δ = 2.0), non-linear change resulted in higher power. Even within the non-linear patterns of NL3 and NL4 which used the same amount of total change (Δ = 1.5), marginally higher power was observed for NL3 (Δ = 0, 1.5) than power in NL4

(Δ = 0.5, 1.0). NL3 implemented change in a step function in which the size of the step function was larger than that in NL4.

Figure 3.23: Mean Power Conditional on Change Patterns



Chapter 4: Discussion

This research study involved conceptualizing and testing the performance of omnibus hypothesis tests of detecting change at an individual level in the context of IRT and CAT. The simulation design was similar to that used by Finkleman et al. (2010) and Lee (2015), but differed with respect to use of multiple occasions, implementation of change in different linear as well as non-linear patterns, and use of omnibus tests to detect change over multiple testing occasions.

Major Effects on Type I error and Power

Statistics

Statistic: Type I Error

The Statistic effect was found to be prominent for Type I error and Power across the change pattern conditions. This was the only effect that contributed to significant amounts of variation in Type I error (Table 3.1a). The Statistic effect contributed to 80.93% of the variation in Type I error. Mean Type I error under different statistic conditions varied from 0.053 to 0.090 (Table 3.1b). Excluding the ST (mean = 0.082) and χ^2_{GD} statistics (mean = 0.090), the mean Type I error for the remaining conditions varied from 0.053 to 0.058. The standard deviations for the statistics varied in the range of 0.002 to 0.019. ST and χ^2_{GD} yielded higher Type I error than other statistic conditions. Their higher mean Type I error was also particularly influenced by high Type I error under the high change conditions. Overall, performance of most statistics was satisfactory in terms of desired Type I error of 0.05.

Statistic: Power

As can be seen in Appendix Table A13, mean power remained between 0.763 to 0.805 across most statistics except for χ^2_{FI}, under which condition the mean power was 0.749. Overall, all the statistics resulted in reasonable power across conditions. There was a trade-off between Type I error and Power, as χ^2_{GD} resulted in highest power followed by ST and then by LR, F2, F1, and χ^2_{FI} (Table A13, Appendix).

Statistic: Power under Linear Change Patterns

Among the linear change pattern conditions, the observed η^2 for power was found to be 1.90%, 5.72% and 28.28%, respectively (Table 3.2a, 3.3a & 3.4a). Mean power for different statistics conditional on change patterns (Table A18 in Appendix) showed increase in mean power as amount of change increased. Mean power ranged between 0.614 to 0.698 for L1, between 0.951 and 0.981 for L2 and between 0.977 and 1.0 for L3. Across all the linear change patterns, χ^2_{GD} resulted in highest power followed by ST, LR, F tests, and lastly by χ^2_{FI}. However, this ordering closely followed the Type I error rates in reverse. The best balance of Type I error and power for linear change was found for LR and F tests.

Statistic: Power under Non-Linear Change Patterns

Observed η^2 was found to be 6.44%, 2.2%, 6.90%, 2.02%, 11.70%, and 33.93% for NL1, NL2, NL3, NL4, NL5, and NL6, respectively. Observed mean power ranged from 0.275 to 0.364 for NL1, from 0.695 to 0.787 for NL2, from 0.883 to 0.994 for NL3, from 0.862 to 0.917 for NL4, from 0.946 to 0.986 for NL5 and from 0.979 to 0.998 for NL6, for various statistics. As for linear change patterns, χ^2_{GD} resulted in highest power followed by ST, LR, F tests, and lastly by χ^2_{FI} under the non-linear change patterns, again reflecting

Type I error in reverse, with the best balance of Type I error and power for LR and F statistics.

For most of the linear as well as non-linear change patterns, the Statistic effect contributed toward significant variation. The observed power under various statistic conditions was influenced by the amount of change. The power across all statistic conditions increased with increase in change (Table A18, Appendix). Within all the change patterns, observed power remained consistent across various statistics.

Agreement Between Statistics

Table 4.1 presents mean proportion agreement between all the statistics across nine θ levels crossed with six bank type conditions. Proportion of agreement was defined as the number of times any two methods would reject or fail to reject the hypothesis of no-change. The obtained statistic under each condition of the simulation design for each examinee was converted into a 1/0 or Yes/No binary result. Then mean proportion of agreement under each condition was calculated by dividing the total number of times any two methods were either scored both 1 or both 0, by the total number of observations under each condition (10,000). Mean proportions under each condition were then averaged across all θ levels, change patterns, and bank conditions to obtain Marginal Mean Agreement between statistics.

Table 4.1 shows that the mean proportion of agreement ranged from 0.94 to 0.99 between all the statistics. F1 agreed 99% of the time with F2 and LR, 96% of the time with ST and χ^2_{GD}, and 94% of the time with χ^2_{FI}. F2 agreed 99% with LR, 97% with ST and χ^2_{GD} and 95% with χ^2_{FI}. LR agreed 97% with ST and χ^2_{GD} and 95% with χ^2_{FI}. ST agreed 99% with χ^2_{GD}, 94% with χ^2_{FI} and χ^2_{FI} agreed 94% with χ^2_{GD}.

Table 4.1: Marginal Mean Agreement Between Statistics Across θ and Bank Type Conditions

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.99	0.96	0.94	0.96
F2		1.00	0.99	0.97	0.95	0.97
LR			1.00	0.97	0.95	0.97
ST				1.00	0.94	0.99
χ^2_{FI}					1.00	0.94
χ^2_{GD}						1.00

All the statistics showed very strong agreement with one another. Statistics following the same distribution (F1 and F2, LR, ST, and χ^2_{GD}) particularly had very high consensus (0.97 to 0.99) with respect to detection of change. χ^2_{FI} showed the least agreement varying from 0.94 to 0.95 with other statistics. This trend can be attributed to χ^2_{FI} showing lower Type I error compared to other statistics. Mean proportions across all the different conditions can be found in the Tables A70 – A88 in the Appendix.

Statistic: Summary

The simulation results indicated that the performance of all the statistics was satisfactory in terms of detecting change. There was a trade-off between Type I error and power. However, all statistics performed well on those two criteria. Considering this trade-off, the LR test seemed to perform the best. Its Type I error remained around 0.05 across most conditions and yielded highest power after χ^2_{GD} and ST test statistics. Although χ^2_{GD} and ST resulted in highest power in all conditions, their Type I error was also high (Table A18, Appendix). Following LR, all the other statistics resulted in mean power very close to that of LR. For these reasons, the best choice for practitioners would be the LR

test, followed by F2 and F1, if the goal is to control for Type I error. If the goal is to identify the maximum number of examinees changed over a period of time without much consideration for incorrect identification, practitioners should consider using χ^2_{GD} or ST test statistics.

The obtained results indicate that the proposed omnibus hypothesis tests to measure individual change performed very well in terms of Type I error as well as power. These hypothesis tests offer an advantage over methods proposed for the two-occasion condition (Finkleman et al., 2010; Lee, 2015): They can be used in two or multi-occasion conditions without inflating Type I error. These methods were derived as an extension of the methods proposed for the two-occasion case (Finkleman et al., 2010; Lee, 2015).

Discrimination

Discrimination: Type I error

With increase in item discrimination, Type I error also increased (Table A11, Appendix). Marginal Type I error varied in the range of 0.0065 to 0.067. However, the Discrimination effect was not substantial in the ANOVA results for Type I error and observed η^2 remained below 5% (Table 3.1a).

Discrimination: Power

A larger effect of discrimination was observed on power. Mean power increased with increase in discrimination, and varied in the range of 0.632 to 0.885.

Discrimination: Power under Linear Change Patterns

The discrimination effect was found to contribute toward a substantial amount of variation in case of L1 and L2 (90.83% and 74.12%, respectively; Table 3.2a and 3.3a). For the L1 change pattern, power varied from 0.381 to 0.892 and from 0.922 to 0.996

across high, medium, and low discrimination conditions. The standard deviation varied in the range of 0.031 to 0.053 for L1 and from 0.0004 to 0.008 for L3. As discrimination increased, power increased across all the linear change patterns.

Discrimination: Power under Non-Linear Change Patterns

The effect of discrimination as reflected in observed η^2 was found to be substantial for all the non-linear change patterns (84.91%, 91.02%, 70.58%, 89.16%, 45.57%, 19.24%, respectively for NL1, NL2, NL3, NL4, NL5 & NL6). As shown in Table A30 in Appendix and Figure 3.4, power increased as discrimination increased.

Under small to medium change patterns (L1, NL1, NL2), large differences in power were observed for different discrimination conditions. However, as the amount of change increased (L2, L3, NL3, NL4, NL5, and NL6), power differences in different discrimination conditions disappeared and medium and high discrimination conditions resulted in high power, ranging from 0.95 to 0.99. These results are in the expected direction and support previous research (Finkleman et al., 2010; Lee, 2015).

Other Minor Effects

θ: θ was found to account for very minor variation in Type I error as observed η^2 remained below 5%. Type I error increased marginally at θ levels below 0. As θ increased above 0, Type I error increased slightly and dropped at θ = 2.0 (Table A14, Appendix). The effect of θ was found to be somewhat substantial for L3, as observed η^2 for θ was found to be 5.18%. In all the other linear and non-linear change patterns, observed η^2 remained less than 5%. With respect to power, it increased slightly at θ levels below 0. However, power decreased at θ = 1.5 and at θ = 2.0. Power decreased as θ levels moved

above 0 (Table A14, Appendix). However, the decrease in power at higher θ levels was due to absence of medium and high change at these higher θ levels.

Bank Information: Information was found to have no substantial effect on Type I error. Type I error was slightly higher under the flat condition than peaked condition (Table A12, Appendix). Peaked item banks, on the other hand, resulted in higher power than flat item banks. Among the linear patterns, peaked item banks resulted in higher power than flat item banks. However, no differences in power were observed for the high linear change pattern of L3 under peaked and flat conditions. For non-linear patterns as well, peaked item banks resulted in slightly higher power than flat item banks (Table A42, Appendix). However, when effect of information on power was investigated conditional on θ, (Table A42, Appendix and Figure 3.17), flat banks resulted in higher power than peaked banks at θ = 2.0 and θ = –2.0. In the middle range of θ, however, peaked banks led to more power than flat banks, likely because they provided more information at those θ levels.

Significant Interactions in ANOVAs

$\theta \times$ *Statistic*: The $\theta \times$ Statistic interaction was found to have a large effect on power for L3, NL5, and NL6 change patterns with observed η^2 of 26.78%, 10.09%, and 11.80%, respectively (Table 3.4a, 3.9a and 3.10a). As can be seen in Figure 3.11, in the case of Type I error, χ^2_{GD} and ST overlapped at θ = –2.0. However, χ^2_{GD} resulted in higher Type I error consistently over other θ levels. χ^2_{FI}, as well, underperformed on Type I error across θ. However, at θ = 1.5 and at θ = 2.0, its Type I error escalated, crossing that of F1, F2, and LR tests. Differences in the effect of statistic on power conditional on θ were not as apparent. However, close inspection revealed that large differences in power at θ = –2.0 between various statistics diminished at higher θ levels. Following the similar trend as for

Type I error, χ^2_{FI} resulted in higher power than that of F1, F2, and LR at $\theta = 2.0$. Similarly, χ^2_{GD} and ST had overlapping power curves at $\theta = -2.0$. But χ^2_{GD} led to slightly higher power at higher θ levels. Thus, most statistics performed consistently across θ levels, with the exception of χ^2_{GD} and χ^2_{FI} which rejected more no-change hypothesis cases at higher θ levels.

$\theta \times$ *Discrimination* × *Statistic*: The three-way interaction of $\theta \times$ Discrimination × Statistic had substantial effect on power for L3, NL5, and NL6 patterns with observed η^2 of 18.20%, 10.01%, and 5.05%, respectively. As can be seen from Appendix Table A69 and Figure 3.13, within all the discrimination conditions, χ^2_{GD} resulted in the highest power followed by ST, LR, F2, F1, and lastly χ^2_{FI} across the θ range. Interaction came into play at extreme θ ends. At $\theta = -2.0$, χ^2_{FI} underperformed in comparison with the other statistics. However, at $\theta = 2.0$, differences between χ^2_{FI} and other statistics diminished and χ^2_{FI} outperformed some of the other statistics. This trend was more apparent in medium and low discrimination conditions than in the high discrimination condition. This result could be attributed to outward bias of MLE in the case of Finkleman et al.'s (2010) z test (Lee, 2015), and therefore may have come into play in the case of χ^2_{FI}, which is derived from Finkleman et al.'s (2010) z test. Figure 3.22 shows that the differences in power of different statistics at $\theta = -2$ diminished as θ moved along the continuum. At $\theta = -2$, χ^2_{FI} underperformed compared to other statistics. However, at $\theta = 2$, the curve representing χ^2_{FI} exceeded that of F1 and overlapped with F2 under the high discrimination condition. This effect was even more profound for medium and low discrimination. In the medium discrimination condition, the power of χ^2_{FI} exceeded that of F1, F2, and LR. In the low discrimination condition, χ^2_{FI} exceeded that of F1, F2, LR, and ST in terms of power. This

interplay was particularly strong for NL6, and did not account for substantial variation in power for other change patterns.

Discrimination × *Information*: The two-way interaction of Discrimination × Information had substantial effect on power in the case of NL6, with observed η^2 of 7.16%. As can be seen from Appendix Table A15, all the bank types resulted in high power for NL6. However, in the high discrimination condition, flat and peaked banks led to similar power. With medium discrimination, flat banks displayed slightly higher power than peaked banks, and in low discrimination, peaked banks displayed more power than flat banks. Close inspection of Appendix Table A54 indicates that in the case of high change (L2, L3, NL3, NL4, and NL5) and high discrimination conditions, flat banks resulted in more power than peaked banks. Conversely, in medium/low discrimination and moderate/small change conditions, peaked banks resulted in more power.

Discrimination × *Information* × *Statistic*: The three-way Discrimination × Information × Statistic interaction had substantial effect on power for the NL6 change pattern, with observed η^2 of 6.90%. As shown in Appendix Table A55, peaked item banks crossed with medium and low discrimination conditions resulted in moderately higher power compared to flat item banks across all statistic conditions. However, when crossed with high discrimination, flat banks led to slightly higher power for F1, F2, ST, and χ^2_{GD} statistics. The Discrimination × Information × Statistic three-way interaction was a sizable effect for NL6, but not for other change conditions in which F1, F2, ST, and χ^2_{GD} resulted in marginally higher power for high-flat item banks.

Overall, performance of all omnibus hypothesis tests remained consistent under various testing conditions. These results indicate that these tests are equally useful in less than ideal real life testing conditions as they are in a perfect setting.

Differences in Linear and Non-Linear Change Patterns

One of the most striking features of the obtained results was that observed power was higher for non-linear change patterns than for linear change patterns for the same magnitude of change. As can be seen from Appendix Table A16 and Figure 3.23, the mean of observed power was the lowest for NL1 (mean power = 0.496), which constituted the least amount of change. This condition consisted of $\Delta = 0, 0.5$ SDs of change across the three occasions and total change of 0.5 SD units. Mean observed power was largest for NL6 (mean power = 0.993). The NL6 condition was a condition of maximum change with $\Delta = 1.0, 1.5$ SD units, with total change of 2.5 SD units change. Observed power increased as amount of change increased.

When linear patterns were compared with non-linear patterns for the same total amount of change, non-linear patterns resulted in higher power than linear patterns. For example, in the L1 vs. NL2 comparison, mean power under L1 was 0.650 and that under NL2 was 0.742. Both these conditions consisted of the same amount of total change of 1.0 SD. In case of L1, this change was introduced gradually over three occasions in a $\Delta = 0.5, 0.5$ pattern. In the case of NL2, this change was introduced over three occasions in step function in $\Delta = 0, 1.0$ pattern. Similarly, in the L2 vs. NL5 comparison, observed mean power for L2 was 0.971 and that for NL5 was 0.974. For both these change patterns, total amount of change at the end of three occasions remained 2.0 SDs. In the case of L2, this change was introduced in a $\Delta = 1.0, 1.0$ change pattern and in the case of NL5, the total

change occurred across three occasions in a Δ = 0.5, 1.5 change pattern. The difference in power was more substantial in the L1 vs. NL2 (total Δ = 1.0) comparison than in the L2 vs. NL5 comparison. The difference in the power resulting from linear or non-linear nature of change diminished as the amount of change increased. This effect on power owing to the nature of change was also observed within the non-linearity of change. The change pattern of NL3 (Δ = 0, 1.5, total Δ = 1.5) yielded observed power of 0.931, whereas the change pattern of NL4 (Δ = 0.5, 1.0, total Δ = 1.5) yielded an observed power of 0.892. Note that both change patterns consisted of the same total amount of change. However, NL3 which introduced this same amount of change in a bigger step function than NL4 resulted in higher power.

In general, the non-linear patterns had a higher effect on power than the linear patterns. This result implies that non-linear change was detected more often than linear change. The reason for this trend could be that when change is non-linear, the difference between any two immediate θs is larger than when the change is gradual or linear. Thus, non-linear patterns had larger effect size than the linear patterns and the larger effect size between the immediate θs for the non-linear pattern resulted in higher power than the linear patterns.

Comparison with Previous Research and Findings

The current research study extended research by Finkleman et al. (2010) and Lee (2015) on measuring intra-individual change in the context of AMC. Finkleman et al. (2010) proposed the hypothesis testing approach for measuring individual change. This approach involves using hypothesis tests for measuring growth when examinee performance is measured using CAT. They proposed hypothesis tests and evaluated their

performance in terms of Type I error and power. The hypothesis tests were designed for the two-occasion case and their performance was tested under various testing conditions. Lee (2015) expanded on their work and proposed two new hypothesis tests in the AMC context, that is, the Score Test and the Kullback-Leibler Divergence Test.

Methodologically, the current study was similar to the previous studies (Finkleman et al., 2010; Lee, 2015). They also used 9 θ levels ranging from –2.0 to 2.0 in 0.5 SD increments. The performance of hypothesis tests was measured in the AMC framework. The parameters used for item banks in the current study were the same as those used by Lee (2015). The amount of change (Δ = 0, 0.5, 1.0, 1.5) was also consistent with the previous work. Lee (2015) also used Δ = 0.25 as a very small change condition. Lee (2015) used item banks consisting of 300 and 500 items in the AMC. A bank of 500 items is considered as adequate for CAT (Thompson & Weiss, 2011); however, Lee (2015) showed that a 300-item AMC bank performed as well as a 500-item AMC bank in terms of Type I error and power.

The current simulation design differed from the earlier work in few ways. One, the Δ = 0.25 condition was excluded from the current study as Lee (2015) found that the observed power was very low, ranging from 0.2 to 0.3 in this condition. The small amount of change was not detected very often and led to incorrect acceptance of the no-change hypothesis most frequently. Second, the previous studies (Finkleman et al., 2010; Lee, 2015) used fixed as well as variable-length conditions. While the present study used only a fixed-length condition, it can be expected that the hypothesis test results would hold in a variable-length condition, at higher efficiency in terms of number of items required to measure individual change. Third, this study used multiple occasions, unlike the previous

studies (Finkleman et al., 2010; Lee, 2015) which used only the two-occasion case to measure individual change. The motivation of the current work was to expand the measurement of change beyond the two-occasion case. Fourth, in conjunction with the generalization of measuring change to multiple occasions, omnibus hypothesis tests were developed to measure change when an examinee is measured at multiple occasions. These hypothesis tests were based on those proposed by Finkleman et al. (2010) and Lee (2015). In addition, two new tests were proposed in the ANOVA framework and their performance was examined. Fifth, previous studies (Finkleman et al., 2010; Lee, 2015) used theoretical Fisher information in calculation of the error term in the denominator of the test. The current study used observed information rather than theoretical information so that the tests would be sensitive to person misfit when used with real data. Using observed information would mirror a real life testing setting, where expected information is unknown. Sixth, Lee (2015) used multiple item selection criteria in CAT. This study used Fisher information as the item selection method, as Lee (2015) did not find substantial differences between various item selection methods as reflected in Type I error and power. Lastly, both Finkleman et al. (2010) and Lee (2015) compared the performance of hypothesis tests for CAT as well as conventional tests. They found that CAT was more efficient in detecting individual change. Only the AMC approach was used in this study to test the performance of the hypothesis tests. However, when IRT-based item parameters have been established, the hypothesis tests can be applied even when conventional tests are used to measure change.

The results found in this study generally agreed with Finkleman et al.'s (2010) and Lee's (2015) results. Type I error for the *Z* test and LR test remained around 0.05. In the

current study, the χ^2 test based on Finkleman et al.'s *Z* test (2010) and the LR test also resulted in Type I error around 0.05. However, in the present simulation, the χ^2 test based on Guo and Drasgow's (2010) *Z* test resulted in Type I error around 0.09. With respect to the Score Test, Lee (2015) found the Type I error of the ST to be around 0.05. However, this test too, resulted in high Type I error, around 0.09 across the θ range in the current study. One possible reason for high Type I error could be the use of observed information instead of theoretical information. In terms of power, the results of this study also generally agreed with the results by Finkleman et al. (2010) and Lee (2015). High discrimination and medium to high change conditions resulted in high power. Results were consistent across the θ range. One apparent trend for Type I error of the statistics across the θ range was that most statistics performed consistently across θ, with an exception of χ^2_{FI}. χ^2_{FI} seemed to underperform at θ = –2.0, and resulted in elevated Type I error at θ = 2.0. This result was also found in Lee's (2015) and Finkleman et al.'s (2010) work, where Type I error of Finkleman's Z test dropped at θ = –2.0 and increased at θ = 2.0. This could be due to bias in MLE. The patterns of bias in MLE have also been reported in previous studies on the behavior of MLE in CAT (Wang & Vispoel, 1998; Wang, Hanson & Lai, 1999; Warm 1989). Lord (1983) derived the bias function of MLE (p. 430) as

$$\text{Bias}(\text{MLE}(\theta)) \simeq \frac{D}{I^2}\sum_{i=1}^{n} a_i I_i(\phi_i - 1/2) \tag{48}$$

where $\phi = (P_i - c_i)/(1 - c_i)$. Equation 48 suggests that the bias will be close to zero if all items have difficulties the same as a simulee's θ: replacing $b_i = \theta$ in Equation 20, P_i becomes $P_i = c_i + \frac{1}{2}(1 - c_i)$ and $\phi = \frac{1}{2}$, which leads to bias of zero. The bias will be negative if θ is smaller than item difficulties: $b_i < \theta$ makes $P_i < c_i + \frac{1}{2}(1 - c_i)$, $\phi < \frac{1}{2}$ and

bias to be negative. The bias will be positive if θ is greater than difficulties (i.e., biased outward). In CAT, this implies that MLE is theoretically unbiased or biased to a small degree as CAT administers items based on each examinee's θ level. CATs in practice, however, will have a slightly larger bias at extreme θ ranges because a sufficient number of items with extreme difficulties may not be available. Equation 48 also implies that bias increases with high discriminating items.

Results showed that χ^2_{FI} had bias in θ estimates in opposite directions; that is, one estimate was biased outward and the other estimate was biased inward. The change estimate $\hat{\theta}_2 - \hat{\theta}_1$ (i.e., the numerator of the *Z* statistic and thereby the χ^2_{FI} statistic) will be larger when θ estimates from the two occasions are biased in opposite directions, which makes the χ^2_{FI} more likely to be rejected. More simulees were rejected at θ = –2.0 and at θ = –2.0 because larger bias in those regions made more $\hat{\theta}_2 - \hat{\theta}_1$ larger, and made it easier for the χ^2_{FI} statistic to be rejected.

Limitations and Future Recommendations

Effect of θ Estimation: Future studies can investigate the effect of θ estimation methods in identifying the significance of individual change. There are four θ estimation techniques that have been primarily investigated in the CAT literature: ML, weighted maximum likelihood, and two Bayesian methods – EAP and MAP. The four θ estimation techniques have shown differences in bias and standard error (SE) in the implementation of CAT (e.g., Wang, Hanson & Lau, 1999; Wang & Vispoel, 1998; Weiss, 1982; Yi et al., 2001). MLE has shown smaller bias than EAP and MAP, and WLE was derived to further reduce the bias in MLE. The Bayesian methods have had the smallest SEs across θ, while MLE had the largest SEs and WLE had SE slightly lower than MLE. Future research is

needed that examines whether θ estimation methods in relation to difference in bias and SE can affect the performance of test statistics and item selection methods in detecting individual change under various item banks.

Higher-Order Interactions in ANOVA: The present study examined two-way or three-way interactions and included other higher order interactions in the error source of variation to keep the error variance under 5%. Further studies can analyze the effect of the higher-order interactions in ANOVA and the effect of combinations of various factors on power for various change patterns over multiple occasions.

Variable-Length Tests: The present study evaluated the performance of the omnibus hypothesis tests for 30-item fixed-length CATs. Previous studies (Finkleman et al., 2010; Lee, 2015) have compared fixed-length CATs with variable-length CATs in measurement and detection of individual change and have found variable-length CATs to be more efficient. The number of items required is significantly less for variable-length CATs compared to fixed-length CATs for the same amount of change (Lee, 2015). Lee (2015) found that using variable-length CAT can help reduce the number of items by as much as 50%. At extreme θ ranges, more items may may be required in the case of variable-length CATs, especially when an item bank is used with other than a flat information function.

Further studies can analyze if the same efficiency is achieved when the new omnibus hypothesis tests for multiple occasions are used to measure change in the AMC framework. It can be expected that variable-length CATs would be more efficient, but further investigation would provide the nature and specifics of the advantages of using

variable-length CATs in terms of the number of item required to identify psychometrically significant change.

Multi-occasions and Evaluation of Post-Hoc Tests: The present study used three occasions to evaluate the performance of omnibus tests. Though the performance should hold even beyond the three-occasion case, empirical evidence should be gathered to test this claim. Once the omnibus test has shown evidence of psychometrically significant change, post-hoc tests (e.g., Finkleman et al.'s Z or LR test, 2010) with adjusted α can be used to determine the location of significant change by performing pairwise comparisons. However, when there are more than three occasions, performance of the post-hoc tests should also be evaluated to establish the location of change.

Generalizations to Other IRT Models: The performance of these hypothesis tests can be studied extending the simulation design from unidimensional/dichotomous IRT models to multidimensional and polytomous IRT models. Although it is common to assume items on a test as measures of a unidimensional latent trait, it is not always justifiable. Many educational and psychological variables have been described as inherently multidimensional, and many personality and health assessments measure multiple dimensions (Ackerman, 1994; Reckase, 1985; Reckase, 2009). Wang and Weiss (2017) have successfully generalized several of the test statistics examined here to the dichotomous multidimensional case for determining significant change on two occasions; these methods are designed to deal with situations where one trait changes and others do not.

The simulations can also be extended to polytomous IRT models. The dichotomous models have major implications in many educational settings. However, it would be

interesting to see whether the omnibus hypothesis tests lead to similar results in the case of polytomous models. Polytomous models are useful in applied psychological measurement or personality testing. Likert-type items are often used in attitude measurement. Similarly, partial credit models are also used in cognitive measurement. Extending the research to various contexts including polytomous and multidimensional models will further the applicability of AMC in the measurement of individual change.

Psychometric Significance vs. Practical Importance: Psychometrically significant individual change may not imply practically or clinically important change across all occasions and vice-versa. For example, psychometrically significant change in a depressed patient may not be enough for a therapist to decide to reduce or stop treatment. Or, students who show psychometrically significant change over multiple occasions may not indicate acquired mastery over the material (assuming achieving mastery is the goal). In such situations, instead of mere psychometric importance, factors like the motivation for the measurement, the nature of the trait being measured, and the expected end result would be considered before a clinician/therapist/trainer/academician can decide to terminate the therapy/treatment/training. In some situations, such as mastery, achievement can be defined by setting a cut-off score. When the examinees obtain the predetermined cut-off score with minimal error, the desired level of learning may be assumed to have occurred. Alternatively, methods such as effect size or confidence interval, taking into account IRT-based error, may also be devised depending on the motivations behind measurement and the goal of learning. Such methods may give more detailed information about the amount as well as the nature of change.

Generalizability of Results: The results obtained in the current simulation study were generated under ideal testing conditions. Consequently, the results should generalize to real life settings when the data are model fitting. However, real life testing conditions might differ from the simulation conditions. In simulations, "true values" of the variables are known and hence it is possible to evaluate performance of the hypothesis tests in terms of variables such as power and Type I error. In real life testing conditions however, there is no way of knowing the true θs. Because there would be no "true" criterion to evaluate change in real data, there would be no means of knowing whether the observed "significant" change actually captured "true" change (e.g., Brouwer et al., 2013). Using the hypothesis tests on real data, it would be possible to determine the proportion of examinees showing significant individual change. However, it would be impossible to know the accuracy of those decisions.

The current study used simulated item banks. To provide a basis for comparison, two test information functions (peaked and flat) were used. In real testing conditions, the "true" item parameters are unknown, and test information is often peaked (Fletcher, 1999; Gibbons et al., 2008; Weiss, 2011). With the peaked test information used in this simulation study, an attempt was made to use an item bank which would be realistic so as to facilitate the interpretation of the results.

Some applied CATs are barely adaptive in practice (Thissen & Mislevy, 2000; Drasgow & Chuah, 2006). Practical constraints such as item exposure controls and content specifications are often imposed. In order to test the performance of the hypothesis tests, constraint free ideal testing conditions were used so that results could be attributed to the manipulated factors in the simulation design. Future studies should evaluate the

performance of the hypothesis tests in detecting change under various conditions reflecting real life CAT applications, including such variables as item parameter estimation error, person misfit, effect of item exposure and content balancing, realistic item banks, and other conditions as deemed critical.

Implications of Results

Overall, the current simulation study established that proposed omnibus tests performed very well in detecting change at an individual level when an individual is measured over multiple occasions. The Type I error also remained around 0.05 for most hypothesis tests. Additionally, the differences in item banks with respect to varying discrimination and type of information did not affect the performance of the statistics substantially. Generally, all omnibus tests performed in high and medium discrimination as well as in flat and peaked item bank conditions. These results imply that the omnibus tests are robust to different testing conditions, which makes them appropriate for use in various real life testing conditions. This study also showed that bank structure was not an important factor in influencing performance of the omnibus tests, so we can conclude that AMC over multiple occasions is reasonably robust to bank structure as long as the bank is large enough to provide information throughout the θ range.

The omnibus hypothesis tests are applicable in an academic, clinical, industrial, or any other type of setting where the emphasis is on learning or change (positive or negative) and quantification of it. The proposed hypothesis tests offer a reasonable way of measuring individual change, including growth and decline. While the hypothesis tests should not be applied in isolation, without understanding or consideration of the practical significance of

change and without expert assessment, the hypothesis tests nevertheless provide a good starting point to assess the psychometric significance of change.

Although simulation conditions generally do not accurately represent real life testing conditions, not all the conditions used in the current simulation were near ideal – for example, peaked item banks with medium or low discriminations were used in the simulations. However, performance of the hypothesis tests was notable in these non-ideal conditions, as well.

Non-linear change patterns were successfully detected more often than linear change patterns. If such non-linear growth occurs in real-life learning situations, the hypothesis tests may be an excellent way of detecting change.

When it comes to implementing the AMC procedure, developing an item bank is instrumental in measuring change adequately. For developing an item bank, size and information need to be determined. Although an item bank of 500 items has been described to be adequate (Thompson & Weiss, 2011), Lee (2015) found that a 300-item bank performed as well as a 500-item bank in measurement of change in terms of Type I error and power. If there are high stakes involved in testing, and security and item exposure are crucial factors, then it may be more useful to develop larger item banks and make sure that there is a sufficient number of items in each content domain.

For measuring growth sufficiently over multiple occasions and changing θ, items should be developed such that they would provide information over a wide range of θ. In practice, item banks may be peaked. However, flat item banks which provide information across the θ continuum would reduce off-target testing and measurement error when the same item bank is used repeatedly. Such item banks can be developed by including items

with varying difficulty and discrimination across θ. At Occasion 2 or later occasions, more difficult items are necessary than easy items since many examinees might be in a higher θ range after positive change (the reverse would also be true if negative change is expected). However, if it is a problem to add difficult items in practice, then the use of AMC needs to make sure to allow enough items to be administered before determining a change decision by setting the appropriate number of items to be administered so that a correct decision of significant change for high θ examinees can be made.

Monte-carlo simulations can be performed with respect to performance of the test statistics once an item bank is developed. Evaluation of the hypothesis tests can be performed under various expected change conditions. Decisions should also be made about the number of items to be administered at each assessment occasion when using a fixed-length test, as measurement of change was affected by the number of items in Lee's (2015) research. An adequate number of items for fixed-length tests can also be found in simulations. For variable-length tests, an adequate number of items can be determined by determining the standard error associated with $\hat{\theta}$, as well as adequate Type I error and power in the monte-carlo simulations.

Chapter 5: Real-Data Analysis

In addition to evaluating the performance of the omnibus hypothesis testing methods using simulations, this study also involved using the omnibus hypothesis tests on real K-12 data. The omnibus tests were used on real K-12 data in order to draw comparisons between the results obtained with the simulated and real data. Such an exercise would not only help in understanding the performance of the methods in simulated data obtained in terms of Type I error and power, but would also help in analyzing results obtained from real data for cross-comparison.

The K-12 data came from a group of students measured using the adaptive measurement of change method. For each examinee, measurements of Math and Reading ability were taken on three occasions – at the beginning of the school year, at the middle of the school year, and at the end of the school year. The omnibus hypothesis tests were applied to the Math data. Item responses of a total of 14,462 students were obtained using CAT at Occasion 1 (early in the school year); 11,585 students were measured at Occasion 2 (midway through the school) and Occasion 1; and 8,979 students were measured at Occasion 3 (the end of the school year), Occasion 2, and Occasion 1. The analyses were based on the latter group. At Occasion 1 and 2, 30-item fixed-length CATs were administered to the students. At Occasion 3, the CAT was terminated after administration of 25 items if the standard error was equal to or less than 0.3. The CAT was terminated after administration of 30 items when the standard error was larger than 0.3. For CATs administered at Occasion 2 and Occasion 3, $\hat{\theta}$s from the previous testing were used at starting points.

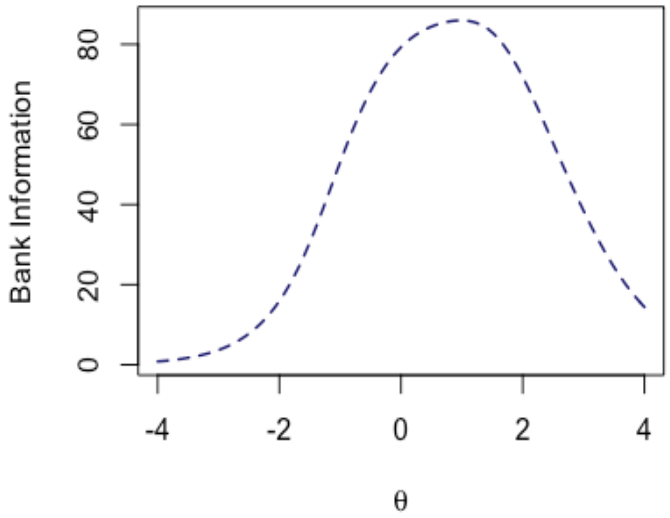
Item Bank

The math item bank consisted of 446 items. Details of the item bank in terms of the item parameter statistics for the 3PL IRT model (with D = 1.7) and test information functions are presented in Table 5.1 and Figure 5.1.

Table 5.1: Mean and Standard Deviation of the Parameters of the Math Item Bank

	a_i	b_i	c_i
Mean	1.32	0.50	0.18
SD	0.38	1.12	0.02

Figure 5.1: Math Bank Information Function



The math item bank was a peaked item bank which peaked around $\theta = 1$. The six omnibus hypothesis tests – F1, F2, LR, χ^2_{FI}, χ^2_{GD} and ST were applied to the K-12 Math data.

Results

Comparison Among Change Detection Methods

Table 5.2 presents the proportion of cases identified as showing psychometrically significant growth by the six omnibus tests.

Table 5.2: Percentage of Examinees with Psychometrically Significant Change for Six Omnibus Tests

F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
22.37%	24.16%	25.31%	27.93%	37.53%	40.90%

Table 5.2 shows that the percentage of examinees with psychometrically significant change varied from 22.37% to 40.90%, showing large variation across the six omnibus hypothesis tests. At 22.37%, the F1 statistic identified the least number of examinees as showing psychometrically significant change. Number of examinees showing significant change as identified by F2 was close to that of F1, at 24.16%. The LR test identified 25.31% examinees to have changed significantly over an academic year. Percentage of examinees identified to be showing significant change was 27.93% for ST. χ^2_{FI} showed 37.53% examinees to be showing significant change while χ^2_{GD} identified 40.90% examinees to have shown significant change in their Math ability.

It is interesting to note that although varied, the omnibus tests that were similar to one another in their formation resulted in proportions which were very close to one another. Thus, the proportion of cases showing significant change were similar for F1 and F2, LR and ST, and χ^2_{FI} and χ^2_{GD}. χ^2_{GD} identified the maximum number of examinees as showing significant growth, while F1 was the most conservative test in identifying change.

Table 5.3: Proportion Agreement Between Omnibus Tests Used on K-12 Data

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{G}
F1	1.00	0.98	0.97	0.94	0.85	0.81
F2		1.00	0.98	0.96	0.87	0.83
LR			1.00	0.97	0.88	0.84
ST				1.00	0.90	0.87
χ^2_{FI}					1.00	0.96
χ^2_{GD}						1.00

Agreement Between Methods

Table 5.3 shows the proportion of agreement between omnibus tests when applied to K-12 data. Any two methods were defined to be in agreement when both either rejected or failed to reject the hypothesis of no-change. Table 5.3 shows very strong agreement between F1, F2, and LR tests with the mean proportion in the range of 0.97 to 0.98. ST also was found to strongly agree with LR and F tests, with the mean agreement ranging from 0.94 to 0.97. χ^2_{FI} showed strongest association with ST followed by LR and F tests. The proportion agreement for χ^2_{FI} ranged from 0.85 to 0.90. The mean agreement for χ^2_{GD} varied in the range of 0.81 to 0.96. χ^2_{GD} agreed most strongly with χ^2_{FI}, followed by ST, LR, and F tests. The F, LR, and ST tests showed very high proportion of agreement. Although χ^2_{FI} and χ^2_{GD} resulted in less agreement compared to other methods, they also displayed reasonable consensus with other methods. χ^2_{GD} had the least agreement with the other methods.

Distribution of Observed Statistics

Figure 5.2 displays distributions of the observed statistics followed by Table 5.4 describing their properties. The mean of F2 was higher than the mean of F1, as evidenced

by the higher detection rate of F2 than F1. Means of the remaining statistics which were distributed as χ^2also increased as per their detection rates. ST and χ^2_{FI} distributions had the largest skew and kurtosis as reflected in Figure 5.2 as well as in Table 5.4. Increase in proportion of cases being detected as showing psychometrically significant change for F1, F2, LR, ST, χ^2_{FI} and χ^2_{GD} statistics, respectively is reflected in increase in mean, SD, skew and kurtosis.

Figure 5.2: Distributions of Observed Statistics in K-12 Math Data

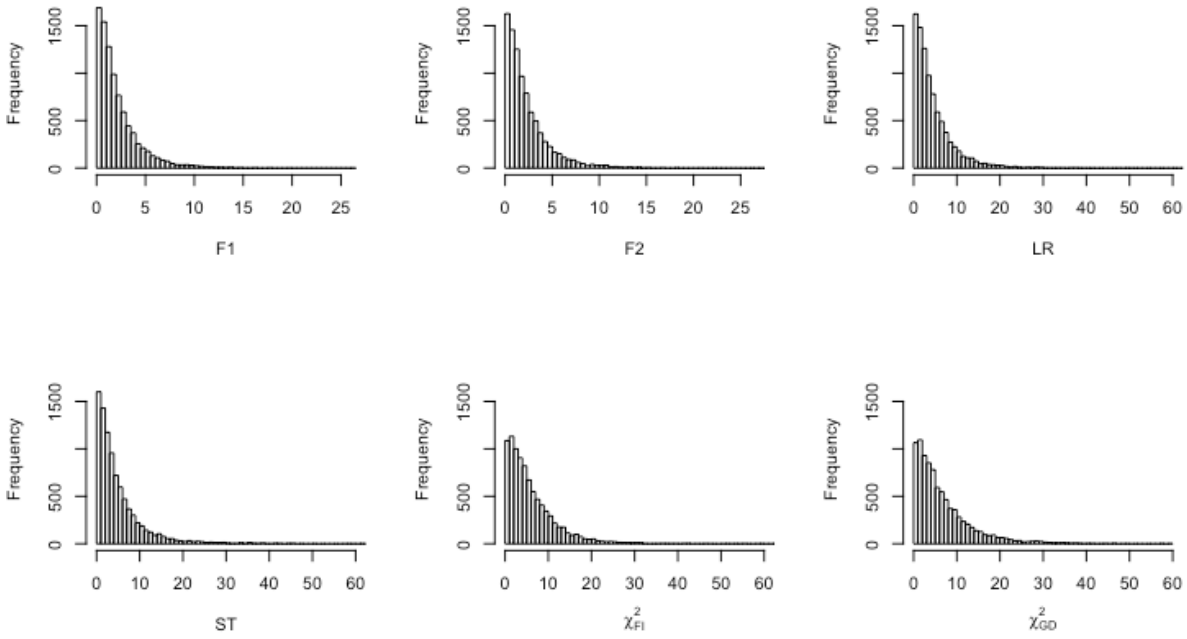


Table 5.4: Descriptive Statistics of Observed Test Statistics on K-12 Data

	Mean	SD	Skew	Kurtosis
F1	2.15	2.17	2.25	7.90
F2	2.25	2.27	2.22	7.66
LR	4.49	4.65	2.76	15.01
ST	5.09	6.20	4.74	49.73
χ^2_{FI}	6.04	5.96	3.36	31.17
χ^2_{GD}	6.67	6.56	2.08	6.40

Distribution of Differences in $\hat{\theta}$s

Figure 5.3 displays frequency distribution of change in $\hat{\theta}$s over multiple occasions plotted in histograms. Figure 5.3a displays the frequency distribution of change in $\hat{\theta}$ from the beginning to the middle of the school year. Figure 5.3b displays change in $\hat{\theta}$ from the middle to the end of the school year and Figure 5.3c displays change in $\hat{\theta}$ from the beginning to the end of the year. One noticeable feature of the histograms is differences in their shape and spread. A very high number of frequencies were observed in Figure 5.3c (around 1,100) above 1 SD of the θ scale. Whereas in Figure 5.3a and Figure 5.3b, the number of observations above 1 SD on θ were around 200 and 300, respectively. A much more dense frequency distribution was observed for change in $\hat{\theta}$ of 1.0 and larger for Figure 5.3c compared to that for Figures 5.3a and 5.3b. Conversely, a more dense distribution was observed between $\hat{\theta}$ difference of 0 to 1 in the case of Figure 5.3a and Figure 5.3b (around 1,000 and 1,400, respectively) compared to Figure 5.3c (around 700). The number of observations of negative growth were also observed to be higher in Figure 5.3a (around 1,200) and Figure 5.3b (around 1,700) compared to Figure 5.3c (around 400). These observations are evidenced by results presented in Table 5.5. As shown in the Table, the distribution of $\hat{\theta}_2 - \hat{\theta}_1$ and $\hat{\theta}_3 - \hat{\theta}_2$ had larger skew and kurtosis compared to the distribution of $\hat{\theta}_3 - \hat{\theta}_1$. The distribution of $\hat{\theta}_3 - \hat{\theta}_2$ as presented in Figure 5.3b, had the highest skew and kurtosis followed by that of $\hat{\theta}_2 - \hat{\theta}_1$ and lastly by $\hat{\theta}_3 - \hat{\theta}_1$. The results show that average change varied from about a quarter of a SD (occasion 2 to occasion 3) to over a half SD from the beginning of the school year to the end. It is also interesting to note changes in θ estimates as high as more than four SDs, and negative change of almost the same magnitude.

Figure 5.3: Distribution of Change in $\hat{\theta}$ Over Multiple Occasions

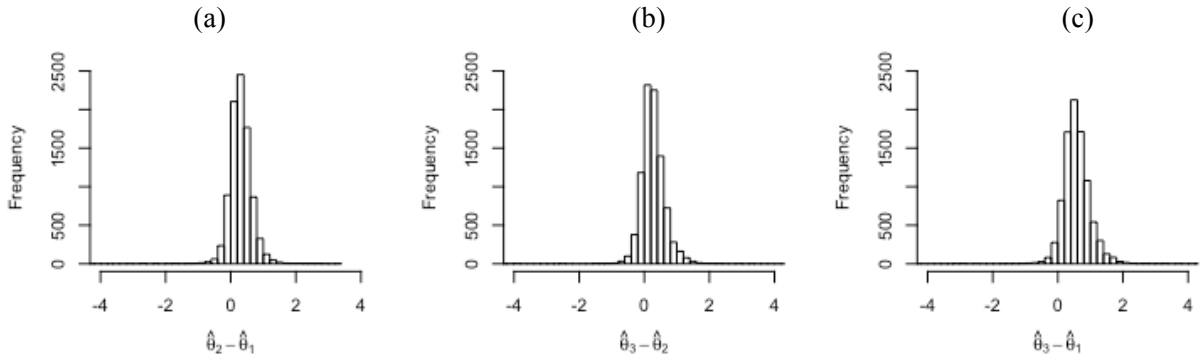


Table 5.5: Descriptive Statistics of Distributions of Change in $\hat{\theta}$

	Mean	SD	Min/Max	Skew	Kurtosis
$\hat{\theta}_2 - \hat{\theta}_1$	0.31	0.35	–3.02/3.27	–1.1	30.31
$\hat{\theta}_3 - \hat{\theta}_2$	0.27	0.37	–3.91/3.24	1.1	45.02
$\hat{\theta}_3 - \hat{\theta}_1$	0.58	0.41	–3.15/4.23	–0.25	16.13

The frequency distribution results and descriptive statistics (Figure 5.3 and Table 5.5) conditional on change in $\hat{\theta}$ across the three measurement occasions indicated that most change tended to occur between occasion 1, the beginning of the school year and occasion 3, the end of the school year.

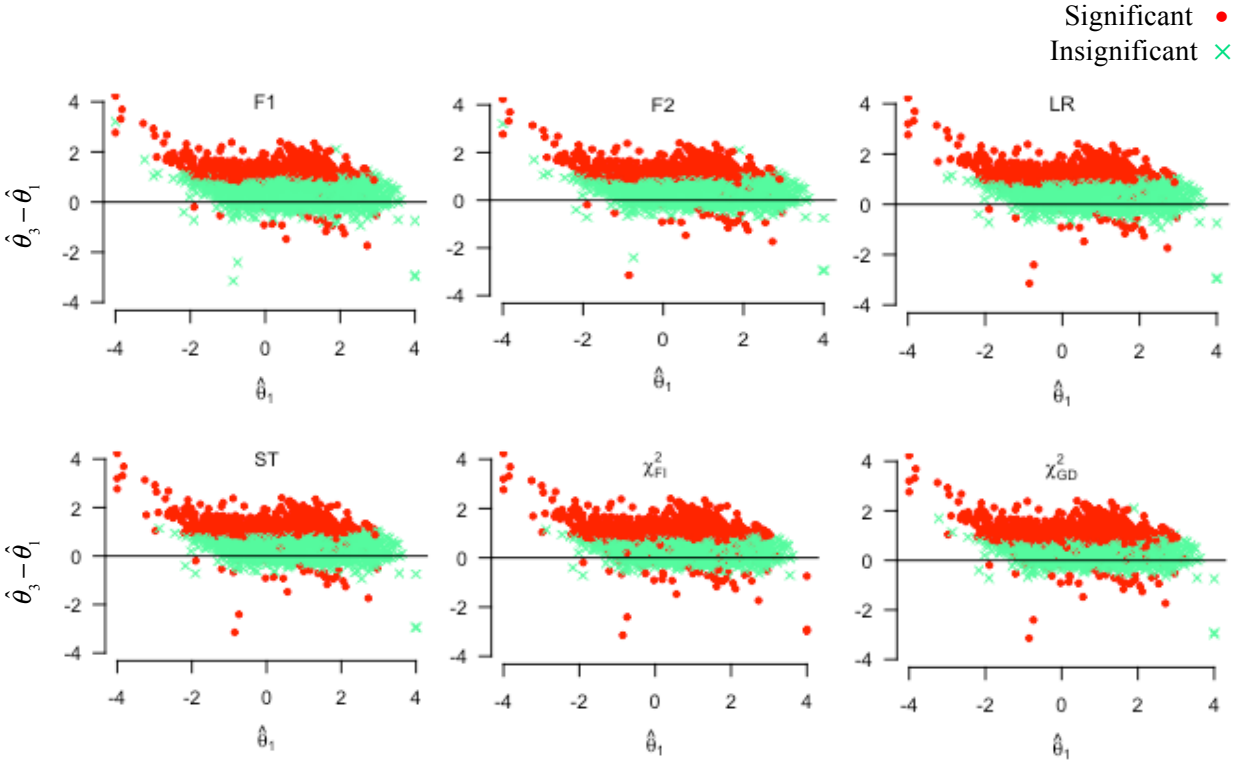
Measured Change as a Function of Initial Status

The location and patterns of observed change across the six omnibus tests were further explored by plotting the difference between $\hat{\theta}$s from occasion 1 and occasion 3 against the initial $\hat{\theta}$ from occasion 1, as shown in Figure 5.4. The figure shows the difference between $\hat{\theta}$s from the beginning and the end of the academic year, plotted against

$\hat{\theta}$ from the beginning of the school year for the six statistics. The cases detected as psychometrically significant by the statistic are presented in red, and the green dots represent the cases detected as non-significant. The black line at a difference of 0 is drawn as a reference for no-change. All six tests showed similar trends in detecting psychometrically significant change. Differences in $\hat{\theta}$s away from 0 were being detected as significant by all tests compared to differences in $\hat{\theta}$s which were close to 0. Results depicted in Table 5.2 are also evidenced in Figure 5.4. F1, F2, LR, ST, χ^2_{FI} and χ^2_{GD} followed that same order in detecting the least to the most cases as significant (Table 5.2). This trend is apparent in Figure 5.4 in the shrinking green belt across the six omnibus tests in that order. It is interesting to note that, except for a small number of students with θ estimates below about –2.5, there was no correlation between the change scores and initial θ estimates, which is a frequent criticism of change scores based in classical test methods (e.g., Cronbach & Furby, 1970; Embretson, 1995; Rogosa & Willett, 1983; Willett, 1994, 1997).

It is important to note that Figure 5.4 presents the $\hat{\theta}$ difference between occasion 1 and occasion 3, whereas the significant and insignificant cases as depicted in the figure have been identified as significant by the omnibus tests. This implies that the change may have occurred at any one or more of the paired intervals. This is the reason why some of the cases close to the difference line at 0.0 appear in red. In the case of these examinees, the significant change may have occurred between occasion 1 and 2 or between occasion 2 and 3, and the change may not have occurred between occasion 1 and 3.

Figure 5.4: Significant vs. Insignificant Cases Across Six Omnibus Tests for $\widehat{\theta}_3 - \widehat{\theta}_1$ Conditional on $\widehat{\theta}_1$

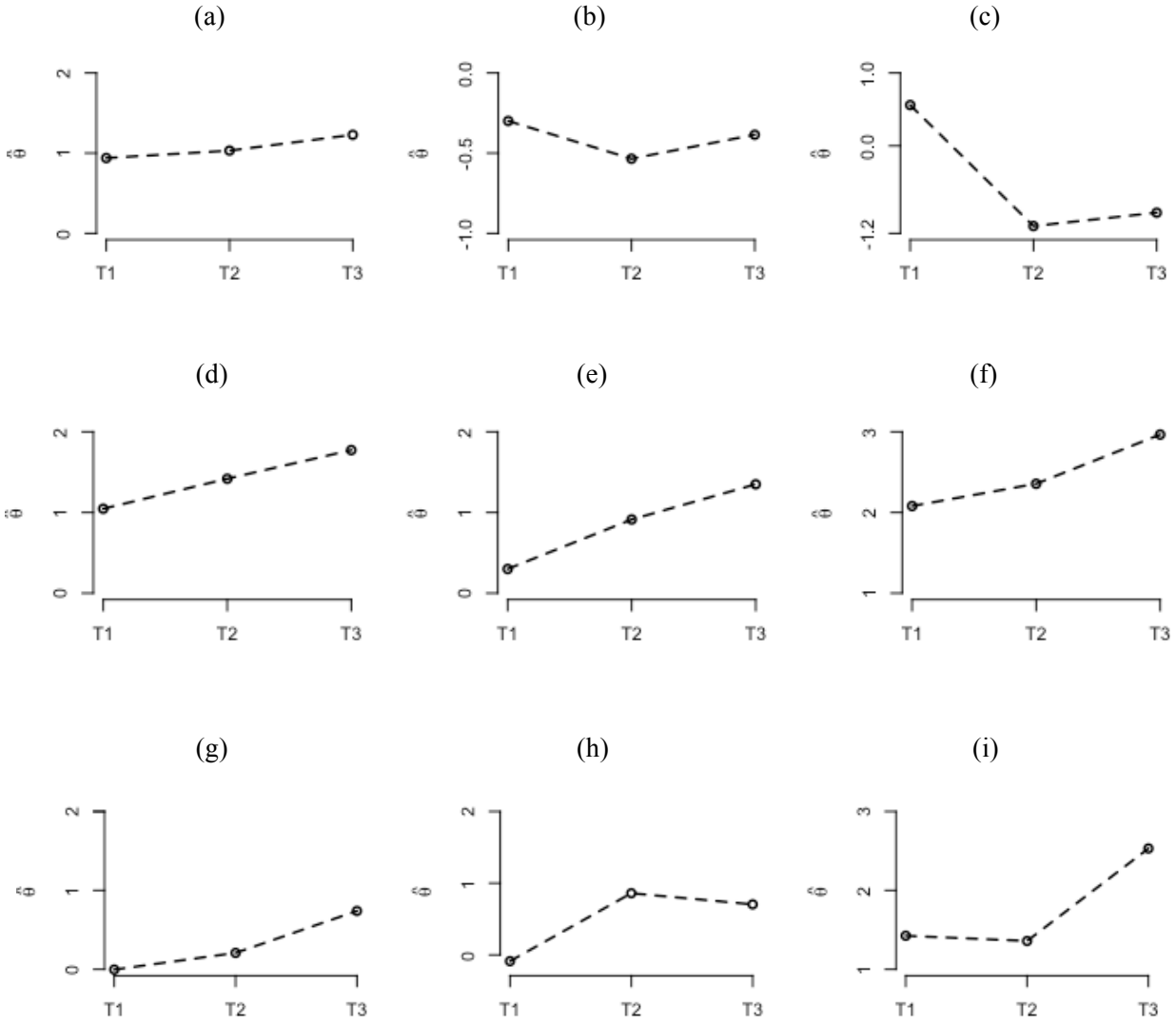


Patterns of Individual Change

Figure 5.5 displays different changing patterns of $\hat{\theta}$ over multiple occasions. Changing $\hat{\theta}$s over an academic year for nine examinees who showed varied change patterns are presented. The examinees were chosen arbitrarily to depict change in their $\hat{\theta}$s.

Figure 5.5 shows that for Examinee (a), there was a slight increase from $\hat{\theta}_1$ to $\hat{\theta}_2$ and from $\hat{\theta}_2$ to $\hat{\theta}_3$ ($\hat{\theta}_1$= 0.94, $\hat{\theta}_2$ = 1.03 and $\hat{\theta}_3$ = 1.23). This change patterns resembled the

Figure 5.5: Changing Patterns of $\hat{\theta}$ over Occasions for Nine Students



"no-change" pattern in the simulations. For Examinee (b), slight negative growth was observed during the first half of the academic year. However, $\hat{\theta}_3$ increased slightly at the end of the year ($\hat{\theta}_1$= –0.30, $\hat{\theta}_2$ = –0.53 and $\hat{\theta}_3$ = –0.39). Such a trend could be explained by motivational factors at occasion 2 or simply by measurement error. Observed change for Examinees (a) and (b) was not detected as psychometrically significant by any of the omnibus tests. In the case of Examinee (c), negative change was observed between the beginning and the middle of the school year, but s/he showed slight positive growth

between the middle and the end of the school year ($\hat{\theta}_1$= 0.56, $\hat{\theta}_2$ = –1.10 and $\hat{\theta}_3$ = –0.92). This negative change pattern was detected to be psychometrically significant by all methods.

Results for examinees (d), (e) and (f) represent observed linear change. For Examinee (d), there was a gradual increase in $\hat{\theta}$ over an academic year ($\hat{\theta}_1$= 1.05, $\hat{\theta}_2$ = 1.42 and $\hat{\theta}_3$ = 1.77). This examinee was detected as showing psychometrically significant change of almost three-fourths of a standard deviation only by the χ^2_{GD} statistic. For Examinee (e) as well, gradual growth in $\hat{\theta}$ occurred over an academic year ($\hat{\theta}_1$= 0.30, $\hat{\theta}_2$ = 0.91 and $\hat{\theta}_3$ = 1.35). The growth of a full standard deviation was detected as psychometrically significant change by all the test statistics. The data also show that Examinee (f) started and ended at a higher level than Examinee (e), even though both demonstrated significant change. Examinee (f) also showed somewhat consistent growth in Math ability. The change appeared to be linear but it was not completely so, as his/her $\hat{\theta}$ changed at slightly different rates across the three occasions ($\hat{\theta}_1$= 2.08, $\hat{\theta}_2$ = 2.36 and $\hat{\theta}_3$ = 2.97). The observed change of almost a full standard deviation was detected as significant by the χ^2_{FI} and χ^2_{GD} statistics.

The data for Examinees (g), (h) and (i) demonstrate non-linear growth. For Examinee (g), more change occurred between the middle and the end of the school year ($\hat{\theta}_1$= –0.004, $\hat{\theta}_2$ = 0.21 and $\hat{\theta}_3$ = 0.74) and the change of almost three-fourths of a SD was detected as significant by χ^2_{FI} and χ^2_{GD} statistics. In the case of Examinee (h), more growth occurred during the first half of the academic year and $\hat{\theta}$ declined slightly at occasion 3 from occasion 2 ($\hat{\theta}_1$ = –0.09, $\hat{\theta}_2$ = 0.86, and $\hat{\theta}_3$ = 0.70). In contrast, Examinee (i) showed most growth during the latter half of the school year compared to the first half ($\hat{\theta}_1$ = 1.43,

$\hat{\theta}_2$ = 1.36, and $\hat{\theta}_3$ = 2.53). Both (h) and (i) examinees were detected as showing significant growth by all the omnibus tests. Although the magnitudes of change and their levels for Examinees (g) and (h) were similar, the manner in which change occurred differed. By contrast, Examinee (i) began at a higher level of θ than the other two, changed more, and had a different level of change.

Comparison with Simulation Results

The TIF for the Math item bank used in the K-12 data was similar to the TIF used for measurement in the simulations – it resembled the high discrimination peaked item bank condition. In terms of the distribution of the parameters, the simulation study used N(1.5, 0.15) for generating the a_i parameter, N(0, 1.2) for generating the b_i parameter and the c_i parameter was kept constant at 0.2. This created the high discrimination peaked (HP) item bank. In the K-12 item bank, the mean and standard deviation of the a_i parameter were 1.32 and 0.38, respectively, those of the b_i parameter were 0.5 and 1.12, respectively and those of the c_i parameter were 0.18 and 0.02, respectively.

In terms of performance of the omnibus hypothesis tests in detecting change, the methods performed reasonably similarly in the K-12 data as in the simulations. The proportions of agreement between the methods were high in both simulation and real data. They most often consensually rejected or failed to reject the hypothesis of no-change in K-12 examinees.

The general order of the hypothesis tests in detecting the proportion of examinees showing psychometrically significant change remained similar in K-12 results to those of the simulation results, with the exception of χ^2_{FI}. In the simulation results, χ^2_{FI} resulted in the lowest power of all the statistics whereas in K-12 data, χ^2_{FI} detected the maximum

number of examinees as showing significant change after χ^2_{GD}. As in the simulation studies, F1 detected the fewest number of examinees as showing significant change, followed by F2, LR, ST and χ^2_{GD}. Another point to note is the high detection rate of χ^2_{FI} and χ^2_{GD} in comparison with other statistics. Although χ^2_{GD} resulted in the highest power in simulations as well (but also had the highest Type I error rate), the differences were more pronounced in the K-12 results than in the simulations. A variety of factors could have contributed to these variations in results of simulation vs. the K-12 data. In particular, the simulations used model-fitting data. In real data, examinee responses may not be ideal. All examinees cannot be assumed to be responding in accordance with the model. Such person-misfit could lead to differences in results. This may also explain the low detection rates for F indexes compared to other tests which are much more dependent on the likelihood function than F indexes. Other factors like the CAT algorithm used in obtaining the data, information structure of the calibrated item banks, item exposure, personality factors of subjects such as a tendency to respond in a specific manner, fatigue, boredom, and lack of motivation could also have attributed to the differences in obtained results under simulation vs. live testing conditions. It would be interesting and also informative to investigate the performance of the omnibus tests under various kinds of misfit. Such analyses would explain the inconsistencies of results between simulated vs. real data.

One of the challenges of using the omnibus methods on real data is that the true magnitude of change which occurred at an individual level in the K-12 data set would not be known. Therefore, on the basis of the simulation results, it is only possible to generalize the results to the K-12 data with reasonable certainty. It is possible to comment on proportion and patterns of change detected across the three testing occasions. However, the

power of the omnibus tests in this K-12 dataset would be unknown as the amount and pattern of true change is unknown.

Overall, the K-12 results indicate that the omnibus hypothesis tests can be applied to real as well as simulated data. The general pattern and detection rates were similar in both sets of data, with the exception of χ^2_{FI}. Considering the high detection rate of the two χ^2_s and less strong agreement with other methods, using LR, ST, or F tests is strongly recommended. The relatively robust performance of the omnibus tests in simulation results under various conditions make them fit to be used on real data.

Conclusions

The present study proposed six omnibus hypothesis tests – F1, F2, LR, ST, χ^2_{FI} and χ^2_{GD} – to identify significant individual change when the measurements are taken over multiple occasions. This research offered significant improvement over previous AMC research (Finkleman et al., 2010; Lee, 2105) for two occasions, as the omnibus hypothesis tests are applicable to the multi-occasion case.

Performance of the omnibus tests was evaluated in terms of Type I error, power and agreement between methods under various testing conditions. All the tests resulted in Type I error in the range of 0.05 to 0.09 and power in the range of 0.7 to 0.8. Observed power was particularly high under high/medium discrimination, flat/peaked item banks and medium/high change conditions. There were no consistent differences in performance of the statistics with respect to various bank types. This confirms that the hypothesis tests are relatively robust under different testing conditions and will be useful in various practical testing situations. In terms of striking a balance between achieving a reasonable Type I error and power, the LR test resulted in a superior performance followed by the F tests, χ^2_{FI}, and lastly by ST and χ^2_{GD}.

The omnibus hypothesis tests have a wide applicability in diverse settings where the focus is on understanding and measuring change in behavior, attitude, clinical symptoms, skills, or ability. Although the current study analyzed the methods in the AMC framework, the methods are equally applicable to conventional tests in testing conditions in which IRT-based item parameters have been established, although their implementation within the CAT framework will result in better and more efficient identification of

significant change due to the superior control of measurement error within CAT. Even though the LR test was found to have the best balance of Type I error and power, the choice of which hypothesis tests to use depends on the purpose of testing. For example, if the practitioners are more focused on identifying change at the cost of making false rejections, ST or χ^2_{GD} might serve them better. Whereas, if it is more beneficial to be conservative instead of making false identification, then χ^2_{FI} may offer the best choice.

The use of hypothesis tests was also demonstrated on K-12 math data. The hypothesis tests were successfully applied to this K-12 CAT data obtained on three occasions over an academic year. The performance of the hypothesis tests was similar in real data as in simulated data. χ^2_{GD} identified most cases to be psychometrically significant followed by ST, LR, F tests and χ^2_{FI}.

These omnibus hypothesis tests should prove to be very useful in K-12 and such educational settings where students are measured over a period of time, e.g., during fall, winte,r and spring semesters and improvement targets are set. The omnibus tests offer a way of evaluating "psychometric significance" of individual change rather than "statistical significance," as the error term in these statistics is rooted in a psychometric framework instead of statistical sampling theory. This approach makes it possible to evaluate individual growth with respect to an examinee's past performance instead of with reference to the group to which s/he belongs. Thus, instead of applying group standards to all students, every student can be evaluated in light of his/her own learning capabilities.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7* (4), 255–278.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Anderson, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis, 95*(1), 1–22.

Arriaga, E.E. (1984). Measuring and explaining the change in life expectancies. *Demography, 21*(1), 83–96.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82,* 60–70.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison, WI: University of Wisconsin Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bock, R. D. (1976). Basic issues in the measurement of change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley.

Bohrnstedt, G. (1969). Observations on the measurement of change. *Sociological Methodology, 1*, 113–133.

Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology, 100*, 192–208.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research, 23* (5), 489–501.

Bryk, A. S., & Randenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newsbury Park, CA: Sage.

Bryk, A. D., & Weisberg, H. I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin, 85*, 950–962.

Burr, J. A., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (vol. 1) (pp. 3–34). Boston: Academic Press.

Cahen, I. S., & and R.L. Linn, R. L. (1971). Regions of significant criterion difference in aptitude-treatment-interaction research. *American Educational Research Journal, 8*, 521–530.

Chandra, T. K. & Joshi, S. N. (1983). Comparison of the likelihood ratio, Wald's and Rao's tests, *Sankhyā, Series A, 45*, 226–246.

Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 12*, 305–308.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical models, design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131–157.

Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18*, 329–337.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical statistics*. New York: Wiley.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thomson Wadsworth.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change" – or should we? *Psychological Bulletin, 74*, 68–80.

De Ayala, R. J., (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–8.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*, 341–353.

Dimitrov, D. M., & Rumrill, Jr., P. D., (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159–165.

Drasgow, F. & Chuah, S. C. (2006). Computer-based testing. In Eid, M. & Diener, E. (Eds). *Handbook of multimethod measurement in psychology* (pp. 87-100). Washington, DC: American Psychological Association.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 184–197). Washington, DC: American Psychological Association.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, *32*, 277–294.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*, 201–212.

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. *Objective Measurement: Theory into Practice, 4*, 223–236.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Falloon, I. H., Boyd, J. L., McGill, C. W., et al. (1985). Family management in the prevention of morbidity of schizophrenia: Clinical outcome of a two-year longitudinal study. *Arch Gen Psychiatry, 42*(9), 887–896.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 403–434.

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*, 238–254.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds). *Advances in psychological and educational measurement* (pp. 97–110). New York: John Wiley & Sons.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3–26.

Fischer, G. H. (2001). Gain scores revisited under an IRT perspective. In A. Boomsma, M. van Duijn, T. Snijders (Eds.) *Essays on item response theory* (pp. 43–68). Springer-Verlag New York, Inc.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement,27*, 3–26.

Fletcher, R. (1999). Incorporating recent advances in measurement in sport and exercise psychology. *Journal of Sport and Exercise Psychology, 21*(1), 24–38.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291.

Fox, J. P., & Glas, A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271–288.

Gagne, D., & Toye, R. C. (1994). The effects of therapeutic touch and relaxation therapy in reducing anxiety. *Archives of Psychiatric Nursing, 8*, 184–189.

Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subset branching strategy in the measurement of classroom achievement* (Research Report 79–6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Available from IACAT (www.iacat.org/biblio).

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D., K., Stover, A., Bock, R. D., Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4*)*, 49–58.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18* (4), 351–364.

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40*(2), 171–178.

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8*, 470–489.

Harris, C. W. (Ed.). (1963). *Problems in the measurement of change*. Madison: University of Wisconsin Press.

Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R., Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59*, 290–298.

Huck, S. W., & McLean, R. A. (1975) Using a repeated measures ANOVA to analyze data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*, 511–518.

Hummel-Rossi, B., & Weinberg, S. L. (1975). Practical guidelines in applying current theories to the measurement of change. I. Problems in measuring change and recommended procedures. *JSAS Catalog of Selected Documents in Psychology, 5*, 226 (Ms. No. 916).

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336–352.

Jennings, E. (1988). Models for pretest-posttest data: Repeated measures ANOVA revisited. *Journal of Educational Statistics, 13*, 273–280.

Johnston, J. & DiNardo, J. (1997) *Econometric methods*. New York, NY: The McGraw-Hill Companies, Inc.

Jöreskog, K. G., & Sörbom, D. (1976). Statistical models and methods for test-retest situations. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley.

Kang, S. M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29*, 87–105.

Kim-Kang, G., & Weiss, D. J. (2007). Comparison of computerized adaptive testing and classical methods for measuring individual change. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Available from IACAT (www.iacat.org/biblio).

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift fur Psychologie, 216*, 49–58.

Kingsbury, G.G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J. Weiss

(Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research, 48*, 370–397.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement, 38*(3), 201–216.

Lee, J. (2015). *Hypothesis testing for adaptive measurement of individual change*. Unpublished Doctoral dissertation, Department of Psychology, University of Minnesota.

Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation*. New York: Springer-Verlag.

Lei, J. & Zhao, Y. (2007). Technology uses and student achievement: A longitudinal study. *Computers & Education, 49*, 284–296.

Li, F., Cohen, A., Bottge, B., & Templin. J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*(2), 181–204.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. Berk (Ed.), *Educational evaluation methodology: The state of the art*. Baltimore, MD: Johns Hopkins University Press.

Linn, L., & Slinde, J.A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47*, 121–150.

Lord, F. M. (1963). The measurement of growth. *Educational and Psychological Measurement, 16*, 421–437.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison: WI: University of Wisconsin Press.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233–245.

Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society, 10,* 888–893.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379–416.

Manning, W. H., & DuBois, P. H. (1962). Correlation methods in research on human learning. *Perceptual and Motor Skills*, 15, 287–321.

Markus, G. (1980). *Models for the analysis of panel data.* Beverly Hills: Sage.

Maurelli, V., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries*. (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of

Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory. Available from IACAT (www.iacat.org/biblio).

Maxwell, S., Delaney, H. D., & J. Manheimer, J. (1985). ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics*, *95*, 136–147.

May, K. & Nicewander, W. A. (1998). Measuring change conventionally and adaptively. *Educational and Psychological Measurement, 58*, 882–897.

McArdle, J. J. (1988). *Dynamic but structural equation modeling of repeated measures data: Handbook of multivariate experimental psychology*. New York, NY: Springer.

McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 57*, 110–133.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*, 126–149.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Mellenbergh G. J., & van den Brink, W. P. (1998). The measurement of individual change. *Psychological Methods, 3*(4), 470–485.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement* 23, 87–89.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20*, 175–240, 263–294.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41,* 582–592.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*, 85–86.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.

Rao, C. R. (1965). *Linear statistical inference and its applications*, New York: Willey.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9* (4), 401–412.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of difference scores in the measurement of change. *Journal of Educational Measurement, 20*, 333–343.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203-228.

Rumrill, Jr., P. D., & Bellini, J. (2009). *Research in rehabilitation counseling*. (2nd ed.). Springfield, IL: Charles C. Thomas.

Schmitt, J. S., & Di Fabio, R. P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology, 57*, 1008–1018.

Shin, J., Espin, C. A., Deno, S. L., & McConnell S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning difficulties. *Asia Pacific Education Review, 5*(2), 136–148.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*, 28–43.

Smits, J. A. J., Berry, A. C., Powers, M. B., Behar, E., & Otto, M. W (2008). Reducing anxiety sensitivity with exercise. *Depression and Anxiety, 25*, 689–699.

Sörbom, D. (1976). A statistical model for the measurement of change in true scores. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measuremen*t. New York: Wiley.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Strenio, J. L. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curves parameters and their relationship to covariates. *Biometrics, 39*, 11–86.

Taniguchi, M. (1988). Asymptotic expansions of the distribution of some statistics for Gaussian ARMA process. *Journal of Multivariate Analysis, 27*, 494–511.

Taniguchi, M. (1991). *Higher Order Asymptotic Theory for Time Series Analysis, Lecture Notes in Statistics, 68,* Springer-Verlag, New York.

Thissen, D., & Mislevy, R.J. (2000). Testing Algorithms. In Wainer, H. (Ed.) *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

Trentacosta, C. J., Criss, M. M., Shaw, D. S., Lacourse, E., Hyde, L. W., & Dishion, T. J. (2011). Antecedents and outcomes of joint trajectories of mother-son

conflict and warmth during middle childhood and adolescence. *Child Development, 82*, 1676–1690.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In van der Linden, W. J., & Glas, C. A. W. (Eds), *Elements of adaptive testing*. (pp. 3–30). New York: Springer.

Von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

Von Eye, A. (1990). *Statistical methods in longitudinal research: Time series and categorical longitudinal data.* Boston, MA: Academic.

Von Minden, S. (2011). *Measuring individual change: A comparison of conventional and adaptive tests.* Unpublished Master's thesis, Department of Psychology, University of Minnesota.

Wang, C. (2014). Reporting reliable change in students' overall and domain abilities across two time points. *Paper presented at the 76th meeting of the National Council on Measurement in Education, Philadelphia, PA .0*

Wang, T., Hanson, B. A. & Lau, C.-M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*, 263–278.

Wang, C., Kohli, N. & Henn, L. (2015). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling, 23* (3), 455–465.

Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, 39, 119–134.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135.

Wang, C. & Weiss, D. J. (2017, in press). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *Journal of the American Statistical Association, 80*, 95–101.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427–450.

Waternaux, C., Laird, N. M., & Ware, J. H. (1985). *Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development* (Tech. Rep.). Cambridge, MA: Harvard University School of Public Health, Department of Biostatistics.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 4*, 473–285.

Weiss, D. J. (2005). *Manual for* POSTSIM*: Posthoc simulation of computerized adaptive testing. Version 2.0*. St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences. 2*(1), 1–23.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Weiss, D. J., & Von Minden, S. (2011). Measuring individual growth with conventional and adaptive tests. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 80–101.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9,* 60–62.

Willett, J.B. (1988-89). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Willett, J. B. (1994). Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671–678). Oxford, UK: Pergamon.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Arnsel & K. A. Reninger (Eds.), *Change and development* (pp. 213-243). Maywah, NJ: Erlbaum.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 59–69.

Wilson, M., Zheng, X., & McGuire, L. W. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement, 13*, 1–22.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*, 50–59.

Yi, Q., Wang, T., & Ban, J-C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38*, 267–292.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154.

Appendix

Table A1: Average Bank Information for Discrimination and Information Conditions

Bank	HD	MD	LD
Peaked	56.515	37.682	20.911
Flat	48.233	32.378	18.674

Table A2 through Table A10: Significant Main Effects, Two-Way and Three-Way Interactions in ANOVAs

Table A2: Mean and SD of Power Conditional on θ for L3 Change Pattern

	θ				
	–2.0	–1.5	–1.0	–0.5	0.0
Mean	0.989	0.996	0.997	0.999	0.999
SD	0.007	0.007	0.005	0.006	0.012

Table A3: Mean and SD of Power Conditional on Statistic and θ for L3 Change Pattern

		θ				
		– 2.0	– 1.5	– 1.0	– 0.5	0
F1	Mean	0.999	0.999	1.000	1.000	0.999
	SD	0.001	0.001	0.001	0.001	0.002
F2	Mean	0.999	0.999	1.000	1.000	0.999
	SD	0.000	0.001	0.001	0.001	0.001
LR	Mean	1.000	1.000	1.000	1.000	0.999
	SD	0.000	0.001	0.001	0.001	0.001
ST	Mean	1.000	1.000	0.999	1.000	0.999
	SD	0.000	0.001	0.001	0.000	0.001
χ^2_{FI}	Mean	0.936	0.975	0.984	0.994	0.996
	SD	0.057	0.011	0.008	0.007	0.005
χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
	SD	0.000	0.000	0.000	0.000	0.001

Table A4: Mean and SD of Power Conditional on Discrimination, Statistic and θ for L3 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
High	F1	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	F2	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	LR	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	ST	Mean	1.000	1.000	0.999	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	χ^2_{FI}	Mean	0.877	0.973	0.980	0.995	0.999
		SD	0.062	0.002	0.005	0.004	0.001
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
Medium	F1	Mean	1.000	1.000	1.000	1.000	0.998
		SD	0.000	0.000	0.000	0.000	0.003
	F2	Mean	1.000	1.000	1.000	1.000	0.998
		SD	0.000	0.000	0.000	0.000	0.003
	LR	Mean	1.000	1.000	1.000	1.000	0.998
		SD	0.000	0.000	0.000	0.000	0.002
	ST	Mean	1.000	1.000	1.000	1.000	0.999
		SD	0.000	0.000	0.000	0.000	0.002
	χ^2_{FI}	Mean	0.945	0.966	0.982	0.998	0.998
		SD	0.022	0.006	0.006	0.002	0.002
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	0.999
		SD	0.000	0.000	0.000	0.000	0.001

Table A4 – Continued on the next page.

Table A4 (continued): Mean and SD of Power Conditional on Discrimination, Statistic and θ for L3 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
Low	F1	Mean	0.999	0.999	0.999	0.999	0.999
		SD	0.000	0.001	0.002	0.001	0.001
	F2	Mean	0.999	0.999	0.999	0.999	0.999
		SD	0.000	0.001	0.002	0.001	0.001
	LR	Mean	0.999	0.999	0.999	0.999	0.999
		SD	0.000	0.001	0.002	0.001	0.001
	ST	Mean	0.999	0.999	0.999	0.999	0.999
		SD	0.000	0.001	0.001	0.000	0.000
	χ^2_{FI}	Mean	0.986	0.987	0.989	0.989	0.992
		SD	0.005	0.012	0.013	0.011	0.008
	χ^2_{GD}	Mean	0.999	0.999	0.999	1.000	0.999
		SD	0.000	0.000	0.001	0.000	0.000

Table A5: Mean and SD of Power Conditional on Statistic and θ for NL5 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
F1	Mean	0.980	0.982	0.981	0.980	0.977	0.974	0.967
	SD	0.030	0.029	0.032	0.035	0.037	0.040	0.048
F2	Mean	0.981	0.983	0.983	0.981	0.979	0.976	0.969
	SD	0.028	0.026	0.029	0.032	0.034	0.037	0.045
LR	Mean	0.982	0.983	0.982	0.980	0.978	0.976	0.970
	SD	0.028	0.027	0.030	0.034	0.037	0.038	0.044
ST	Mean	0.986	0.984	0.968	0.984	0.984	0.981	0.965
	SD	0.019	0.017	0.032	0.024	0.025	0.028	0.037
χ^2_{FI}	Mean	0.876	0.947	0.963	0.958	0.954	0.961	0.965
	SD	0.082	0.031	0.042	0.059	0.066	0.048	0.041
χ^2_{GD}	Mean	0.989	0.990	0.989	0.988	0.986	0.984	0.979
	SD	0.018	0.017	0.018	0.021	0.022	0.025	0.030

Table A6: Mean and SD of Power Conditional on Discrimination, Statistic and θ for NL5 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
High	F1	Mean	0.999	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	F2	Mean	0.999	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	LR	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	ST	Mean	0.998	0.989	0.942	0.995	1.000
		SD	0.000	0.001	0.039	0.006	0.000
	χ^2_{FI}	Mean	0.785	0.958	0.995	0.994	0.997
		SD	0.084	0.011	0.003	0.006	0.001
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
Medium	F1	Mean	0.999	0.999	1.000	0.999	0.999
		SD	0.001	0.000	0.000	0.000	0.000
	F2	Mean	0.999	0.999	1.000	0.999	0.999
		SD	0.000	0.000	0.000	0.000	0.000
	LR	Mean	0.999	0.999	1.000	0.999	0.999
		SD	0.001	0.000	0.000	0.000	0.001
	ST	Mean	0.999	0.998	0.999	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	χ^2_{FI}	Mean	0.943	0.970	0.979	0.979	0.972
		SD	0.011	0.007	0.003	0.007	0.025
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000

Table A6 – Continued on the next page.

Table A6 (continued): Mean and SD of Power Conditional on Discrimination, Statistic and θ for NL5 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
Low	F1	Mean	0.941	0.946	0.944	0.940	0.932
		SD	0.009	0.009	0.029	0.035	0.031
	F2	Mean	0.946	0.951	0.949	0.944	0.938
		SD	0.010	0.010	0.026	0.033	0.028
	LR	Mean	0.946	0.951	0.947	0.941	0.934
		SD	0.014	0.014	0.029	0.037	0.034
	ST	Mean	0.963	0.966	0.964	0.958	0.954
		SD	0.011	0.011	0.020	0.027	0.019
	χ^2_{FI}	Mean	0.899	0.913	0.914	0.902	0.894
		SD	0.009	0.009	0.039	0.087	0.098
	χ^2_{GD}	Mean	0.966	0.969	0.968	0.965	0.960
		SD	0.007	0.007	0.017	0.023	0.018

Table A7: Mean and SD of Power Conditional on Statistic and θ for NL6 Change Pattern

		θ					
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5
F1	Mean	0.996	0.996	0.996	0.996	0.996	0.994
	SD	0.006	0.006	0.006	0.007	0.008	0.010
F2	Mean	0.996	0.997	0.997	0.997	0.996	0.994
	SD	0.006	0.005	0.006	0.006	0.007	0.009
LR	Mean	0.997	0.997	0.997	0.996	0.996	0.995
	SD	0.006	0.005	0.006	0.007	0.007	0.008
ST	Mean	0.997	0.996	0.998	0.998	0.997	0.996
	SD	0.004	0.003	0.004	0.004	0.005	0.006
χ^2_{FI}	Mean	0.960	0.975	0.980	0.980	0.989	0.990
	SD	0.012	0.013	0.017	0.025	0.021	0.015
χ^2_{GD}	Mean	0.998	0.998	0.998	0.998	0.998	0.997
	SD	0.004	0.003	0.003	0.004	0.004	0.005

Table A8: Mean and SD of Power Conditional on Discrimination and Information for NL6 Change Pattern

		HD	MD	LD
Flat	Mean	0.997	0.998	0.980
	SD	0.003	0.001	0.003
Peaked	Mean	0.997	0.997	0.993
	SD	0.003	0.004	0.003

Table A9: Mean and SD of Power Conditional on Discrimination, Statistic and θ for NL6 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
High	F1	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	F2	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	LR	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	ST	Mean	0.999	0.995	1.000	1.000	1.000
		SD	0.000	0.002	0.000	0.000	0.000
	χ^2_{FI}	Mean	0.955	0.982	0.984	0.981	0.998
		SD	0.004	0.005	0.007	0.004	0.000
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
Medium	F1	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	F2	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	LR	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	ST	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000
	χ^2_{FI}	Mean	0.958	0.971	0.988	0.997	0.998
		SD	0.023	0.021	0.003	0.001	0.003
	χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
		SD	0.000	0.000	0.000	0.000	0.000

Table A9 – Continued on next page.

Table A9 (continued): Mean and SD of Power Conditional on Discrimination, Statistic and θ for NL6 Change Pattern

			θ				
			– 2.0	– 1.5	– 1.0	– 0.5	0
Low	F1	Mean	0.989	0.990	0.990	0.989	0.987
		SD	0.005	0.005	0.007	0.008	0.009
	F2	Mean	0.990	0.991	0.991	0.990	0.988
		SD	0.005	0.004	0.007	0.007	0.007
	LR	Mean	0.990	0.990	0.990	0.990	0.988
		SD	0.006	0.005	0.008	0.008	0.008
	ST	Mean	0.993	0.994	0.993	0.993	0.991
		SD	0.005	0.003	0.005	0.005	0.004
	χ^2_{FI}	Mean	0.966	0.971	0.968	0.962	0.970
		SD	0.008	0.013	0.031	0.044	0.035
	χ^2_{GD}	Mean	0.994	0.994	0.995	0.994	0.993
		SD	0.004	0.003	0.004	0.005	0.005

Table A10: Mean and SD of Power Conditional on Discrimination, Information and Statistic for NL6 Change Pattern

		HF	HP	MF	MP	LF	LP
F1	Mean	1.000	0.989	1.000	1.000	0.983	0.992
	SD	0.000	0.026	0.000	0.000	0.003	0.004
F2	Mean	1.000	0.989	1.000	1.000	0.985	0.993
	SD	0.000	0.025	0.000	0.000	0.003	0.004
LR	Mean	1.000	0.989	1.000	1.000	0.984	0.993
	SD	0.000	0.026	0.000	0.000	0.002	0.003
ST	Mean	0.999	0.990	1.000	1.000	0.989	0.995
	SD	0.001	0.021	0.000	0.000	0.001	0.003
χ^2_{FI}	Mean	0.983	0.970	0.989	0.980	0.951	0.987
	SD	0.015	0.029	0.008	0.025	0.012	0.009
χ^2_{GD}	Mean	1.000	0.992	1.000	1.000	0.991	0.996
	SD	0.000	0.020	0.000	0.000	0.001	0.002

Table A11 Through Table A69: Marginal and Conditional Results

Table A11: Mean and SD of Type I error and Power Conditional on Discrimination

		HD	MD	LD
Type I Error	Mean	0.067	0.066	0.065
	SD	0.002	0.002	0.003
Power	Mean	0.885	0.798	0.632
	SD	0.081	0.116	0.152

Table A12: Mean and SD of Type I error and Power Conditional on Information

		Flat	Peaked
Type I Error	Mean	0.068	0.064
	SD	0.002	0.002
Power	Mean	0.765	0.778
	SD	0.101	0.129

Table A13: Mean and SD of Type I error and Power Conditional on Statistic

		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
Type I Error	Mean	0.053	0.058	0.053	0.082	0.055	0.090
	SD	0.002	0.002	0.019	0.006	0.005	0.003
Power	Mean	0.763	0.771	0.773	0.793	0.749	0.805
	SD	0.126	0.123	0.118	0.118	0.101	0.111

Table A14: Mean and SD of Type I error and Power Conditional on θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Type I Error	Mean	0.069	0.067	0.067	0.065	0.066	0.067	0.065	0.066	0.063
	SD	0.018	0.017	0.016	0.014	0.015	0.015	0.017	0.016	0.015
Power	Mean	0.816	0.833	0.842	0.842	0.839	0.705	0.779	0.684	0.495
	SD	0.239	0.231	0.223	0.221	0.222	0.297	0.237	0.244	0.203

Table A15: Mean and SD of Type I error and Power Conditional on Bank Type

		HF	HP	MF	MP	MF	MP
Type I Error	Mean	0.070	0.065	0.068	0.065	0.067	0.063
	SD	0.003	0.003	0.004	0.003	0.003	0.003
Power	Mean	0.895	0.886	0.791	0.806	0.621	0.643
	SD	0.073	0.098	0.108	0.125	0.134	0.170

Table A16: Mean and SD of Power Conditional on Change Pattern

	L1	L2	L3	NL1	NL2	NL3	NL4	NL5	NL6
Mean	0.650	0.971	0.998	0.307	0.742	0.931	0.892	0.974	0.994
SD	0.043	0.005	0.001	0.023	0.034	0.011	0.016	0.005	0.002

Table A17: Mean and SD of Power Conditional on Change Pattern and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
L1	Mean	0.630	0.663	0.687	0.687	0.680	0.669	0.652	0.625	0.554
	SD	0.209	0.216	0.217	0.218	0.227	0.235	0.231	0.220	0.201
L2	Mean	0.968	0.975	0.975	0.973	0.972	0.969	0.962	-	-
	SD	0.038	0.033	0.035	0.040	0.043	0.045	0.049	-	-
L3	Mean	0.989	0.996	0.997	0.999	0.999	-	-	-	-
	SD	0.032	0.010	0.007	0.003	0.002	-	-	-	-
NL1	Mean	0.274	0.302	0.320	0.329	0.327	0.330	0.314	0.298	0.269
	SD	0.113	0.129	0.143	0.144	0.145	0.162	0.152	0.131	0.114
NL2	Mean	0.722	0.754	0.773	0.771	0.765	0.755	0.745	0.725	0.664
	SD	0.198	0.195	0.197	0.195	0.204	0.213	0.215	0.213	0.204
NL3	Mean	0.915	0.936	0.942	0.940	0.937	0.932	0.929	0.914	-
	SD	0.101	0.074	0.070	0.074	0.085	0.088	0.089	0.102	-
NL4	Mean	0.887	0.903	0.908	0.906	0.898	0.891	0.883	0.859	-
	SD	0.115	0.114	0.112	0.117	0.125	0.131	0.136	0.147	-
NL5	Mean	0.966	0.978	0.978	0.979	0.977	0.975	0.969	-	-
	SD	0.056	0.027	0.031	0.035	0.038	0.035	0.039	-	-
NL6	Mean	0.991	0.993	0.994	0.994	0.995	0.994	-	-	-
	SD	0.015	0.010	0.010	0.012	0.010	0.009	-	-	-

Table A18: Mean and SD of Power Conditional on Change Pattern and Statistic

		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
L1	Mean	0.622	0.636	0.642	0.680	0.614	0.698
	SD	0.051	0.050	0.040	0.046	0.032	0.044
L2	Mean	0.969	0.972	0.973	0.978	0.951	0.981
	SD	0.007	0.007	0.005	0.006	0.012	0.005
L3	Mean	0.999	0.999	1.000	1.000	0.977	1.000
	SD	0.000	0.000	0.000	0.000	0.024	0.000
NL1	Mean	0.277	0.291	0.290	0.347	0.275	0.364
	SD	0.025	0.025	0.021	0.029	0.021	0.027
NL2	Mean	0.724	0.736	0.735	0.772	0.695	0.787
	SD	0.040	0.038	0.035	0.043	0.028	0.036
NL3	Mean	0.932	0.936	0.935	0.942	0.883	0.954
	SD	0.012	0.011	0.011	0.017	0.045	0.010
NL4	Mean	0.883	0.890	0.891	0.908	0.862	0.917
	SD	0.020	0.019	0.016	0.020	0.020	0.017
NL5	Mean	0.977	0.979	0.979	0.979	0.946	0.986
	SD	0.005	0.005	0.005	0.009	0.032	0.004
NL6	Mean	0.996	0.996	0.996	0.997	0.979	0.998
	SD	0.001	0.001	0.001	0.001	0.011	0.000

Table A19: Mean and SD of Type I error Conditional on Statistic and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
F1	Mean	0.054	0.053	0.054	0.053	0.054	0.055	0.053	0.052	0.049
	SD	0.008	0.002	0.004	0.003	0.003	0.005	0.006	0.006	0.006
F2	Mean	0.060	0.058	0.059	0.058	0.059	0.061	0.058	0.058	0.054
	SD	0.063	0.059	0.059	0.058	0.059	0.060	0.056	0.058	0.056
LR	Mean	0.063	0.059	0.059	0.058	0.059	0.060	0.056	0.058	0.056
	SD	0.091	0.088	0.086	0.080	0.082	0.080	0.080	0.079	0.071
ST	Mean	0.091	0.088	0.086	0.080	0.082	0.080	0.080	0.079	0.071
	SD	0.055	0.051	0.052	0.054	0.054	0.057	0.051	0.059	0.066
χ^2_{FI}	Mean	0.055	0.051	0.052	0.054	0.054	0.057	0.051	0.059	0.066
	SD	0.090	0.091	0.092	0.088	0.091	0.092	0.091	0.091	0.083
χ^2_{GD}	Mean	0.090	0.091	0.092	0.088	0.091	0.092	0.091	0.091	0.083
	SD	0.012	0.003	0.004	0.004	0.004	0.005	0.007	0.009	0.009

Table A20: Mean and SD of Power Conditional on Statistic and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
F1	Mean	0.812	0.827	0.834	0.835	0.831	0.805	0.803	0.662	0.458
	SD	0.098	0.097	0.099	0.101	0.105	0.125	0.126	0.183	0.190
F2	Mean	0.819	0.833	0.840	0.840	0.837	0.811	0.810	0.673	0.473
	SD	0.094	0.094	0.096	0.098	0.102	0.122	0.123	0.179	0.189
LR	Mean	0.822	0.834	0.840	0.839	0.835	0.809	0.810	0.677	0.487
	SD	0.090	0.093	0.096	0.098	0.103	0.122	0.119	0.175	0.181
ST	Mean	0.847	0.856	0.857	0.856	0.856	0.833	0.826	0.696	0.509
	SD	0.083	0.084	0.083	0.087	0.093	0.111	0.115	0.167	0.188
χ^2_{FI}	Mean	0.746	0.791	0.812	0.814	0.811	0.785	0.799	0.677	0.508
	SD	0.049	0.077	0.091	0.093	0.106	0.123	0.103	0.152	0.159
χ^2_{GD}	Mean	0.849	0.860	0.866	0.867	0.864	0.843	0.841	0.720	0.537
	SD	0.082	0.081	0.083	0.085	0.088	0.105	0.107	0.161	0.184

Table A21: Mean and SD of Power Conditional on Statistic and θ for L1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
F1	Mean	0.608	0.634	0.664	0.666	0.658	0.648	0.626	0.592	0.505
	SD	0.238	0.274	0.246	0.249	0.256	0.264	0.259	0.249	0.227
F2	Mean	0.621	0.647	0.677	0.678	0.671	0.660	0.639	0.606	0.521
	SD	0.235	0.269	0.241	0.244	0.251	0.259	0.255	0.245	0.225
LR	Mean	0.628	0.650	0.680	0.680	0.673	0.660	0.641	0.617	0.553
	SD	0.229	0.265	0.239	0.242	0.250	0.258	0.252	0.239	0.213
ST	Mean	0.678	0.693	0.719	0.717	0.711	0.702	0.679	0.645	0.573
	SD	0.219	0.249	0.224	0.230	0.236	0.242	0.242	0.237	0.218
χ^2_{FI}	Mean	0.559	0.595	0.645	0.644	0.640	0.626	0.625	0.620	0.573
	SD	0.185	0.230	0.223	0.214	0.238	0.252	0.242	0.213	0.184
χ^2_{GD}	Mean	0.687	0.706	0.735	0.736	0.729	0.721	0.700	0.671	0.596
	SD	0.215	0.241	0.216	0.219	0.224	0.231	0.233	0.228	0.215

Table A22: Mean and SD of Power Conditional on Statistic and θ for L2 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
F1	Mean	0.972	0.975	0.974	0.973	0.971	0.965	0.955
	SD	0.041	0.039	0.042	0.044	0.047	0.054	0.063
F2	Mean	0.974	0.977	0.977	0.975	0.973	0.967	0.958
	SD	0.038	0.036	0.038	0.041	0.044	0.050	0.058
LR	Mean	0.975	0.978	0.977	0.975	0.973	0.969	0.963
	SD	0.038	0.035	0.038	0.041	0.044	0.048	0.054
ST	Mean	0.981	0.983	0.982	0.981	0.979	0.975	0.967
	SD	0.029	0.027	0.030	0.032	0.033	0.038	0.047
χ^2_{FI}	Mean	0.925	0.951	0.958	0.949	0.956	0.958	0.960
	SD	0.032	0.037	0.042	0.055	0.068	0.059	0.048
χ^2_{GD}	Mean	0.983	0.985	0.985	0.984	0.982	0.979	0.971
	SD	0.026	0.024	0.025	0.027	0.030	0.033	0.040

Table A23: Mean and SD of Power Conditional on Statistic and θ for L3 Change Pattern

		θ				
		– 2.0	– 1.5	– 1.0	– 0.5	0
F1	Mean	0.999	0.999	1.000	1.000	0.999
	SD	0.001	0.001	0.001	0.001	0.002
F2	Mean	0.999	0.999	1.000	1.000	0.999
	SD	0.000	0.001	0.001	0.001	0.001
LR	Mean	1.000	1.000	1.000	1.000	0.999
	SD	0.000	0.001	0.001	0.001	0.001
ST	Mean	1.000	1.000	0.999	1.000	0.999
	SD	0.000	0.001	0.001	0.000	0.001
χ^2_{FI}	Mean	0.936	0.975	0.984	0.994	0.996
	SD	0.057	0.011	0.008	0.007	0.005
χ^2_{GD}	Mean	1.000	1.000	1.000	1.000	1.000
	SD	0.000	0.000	0.000	0.000	0.001

Table A24: Mean and SD of Power Conditional on Statistic and θ for NL1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
F1	Mean	0.240	0.270	0.290	0.300	0.299	0.302	0.287	0.268	0.234
	SD	0.109	0.126	0.146	0.149	0.148	0.165	0.156	0.134	0.112
F2	Mean	0.254	0.284	0.305	0.314	0.312	0.316	0.300	0.282	0.249
	SD	0.111	0.130	0.148	0.153	0.151	0.168	0.159	0.137	0.116
LR	Mean	0.263	0.288	0.303	0.311	0.308	0.311	0.295	0.280	0.255
	SD	0.107	0.129	0.144	0.147	0.149	0.168	0.155	0.131	0.113
ST	Mean	0.332	0.355	0.370	0.369	0.369	0.368	0.350	0.324	0.286
	SD	0.130	0.148	0.165	0.164	0.168	0.180	0.165	0.145	0.131
χ^2_{FI}	Mean	0.228	0.254	0.271	0.294	0.286	0.296	0.280	0.283	0.279
	SD	0.080	0.108	0.115	0.127	0.123	0.164	0.156	0.126	0.112
χ^2_{GD}	Mean	0.330	0.361	0.380	0.388	0.386	0.390	0.373	0.351	0.313
	SD	0.133	0.145	0.162	0.164	0.162	0.181	0.167	0.149	0.130

Table A25: Mean and SD of Power Conditional on Statistic and θ for NL2 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
F1	Mean	0.704	0.738	0.757	0.758	0.752	0.744	0.729	0.703	0.635
	SD	0.237	0.229	0.229	0.229	0.234	0.242	0.247	0.248	0.236
F2	Mean	0.716	0.750	0.768	0.768	0.762	0.754	0.740	0.716	0.650
	SD	0.230	0.223	0.223	0.223	0.228	0.236	0.241	0.241	0.231
LR	Mean	0.725	0.751	0.767	0.764	0.756	0.746	0.735	0.715	0.654
	SD	0.216	0.218	0.222	0.222	0.231	0.240	0.240	0.236	0.223
ST	Mean	0.776	0.799	0.807	0.801	0.795	0.788	0.773	0.742	0.669
	SD	0.193	0.195	0.195	0.196	0.208	0.215	0.218	0.223	0.220
χ^2_{FI}	Mean	0.635	0.679	0.723	0.717	0.717	0.697	0.706	0.709	0.672
	SD	0.158	0.173	0.200	0.183	0.210	0.223	0.226	0.205	0.190
χ^2_{GD}	Mean	0.777	0.804	0.816	0.816	0.810	0.801	0.788	0.766	0.702
	SD	0.197	0.189	0.190	0.190	0.195	0.204	0.209	0.214	0.213

Table A26: Mean and SD of Power Conditional on Statistic and θ for NL3 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
F1	Mean	0.929	0.940	0.943	0.942	0.937	0.931	0.924	0.907
	SD	0.095	0.085	0.085	0.088	0.094	0.099	0.107	0.126
F2	Mean	0.935	0.944	0.947	0.946	0.941	0.936	0.928	0.913
	SD	0.088	0.080	0.080	0.083	0.089	0.094	0.102	0.120
LR	Mean	0.939	0.944	0.945	0.944	0.938	0.932	0.927	0.914
	SD	0.081	0.080	0.082	0.086	0.094	0.099	0.101	0.116
ST	Mean	0.954	0.957	0.950	0.936	0.952	0.948	0.936	0.905
	SD	0.061	0.059	0.054	0.061	0.073	0.077	0.081	0.103
χ^2_{FI}	Mean	0.778	0.867	0.905	0.907	0.895	0.893	0.910	0.910
	SD	0.114	0.065	0.078	0.084	0.110	0.110	0.097	0.094
χ^2_{GD}	Mean	0.956	0.961	0.963	0.962	0.958	0.953	0.947	0.934
	SD	0.061	0.057	0.057	0.060	0.065	0.070	0.076	0.093

Table A27: Mean and SD of Power Conditional on Statistic and θ for NL4 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
F1	Mean	0.883	0.896	0.900	0.900	0.892	0.883	0.871	0.841
	SD	0.141	0.135	0.134	0.137	0.145	0.153	0.165	0.177
F2	Mean	0.890	0.902	0.906	0.905	0.898	0.889	0.878	0.849
	SD	0.133	0.128	0.127	0.131	0.138	0.146	0.157	0.171
LR	Mean	0.894	0.902	0.906	0.904	0.895	0.887	0.879	0.858
	SD	0.129	0.128	0.127	0.132	0.141	0.148	0.154	0.166
ST	Mean	0.916	0.923	0.923	0.921	0.916	0.907	0.894	0.866
	SD	0.105	0.104	0.103	0.111	0.116	0.123	0.135	0.156
χ^2_{FI}	Mean	0.818	0.865	0.881	0.879	0.862	0.861	0.871	0.861
	SD	0.096	0.126	0.125	0.132	0.149	0.154	0.141	0.135
χ^2_{GD}	Mean	0.920	0.928	0.931	0.930	0.924	0.916	0.907	0.881
	SD	0.102	0.098	0.097	0.100	0.107	0.113	0.124	0.140

Table A28: Mean and SD of Power Conditional on Statistic and θ for NL5 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
F1	Mean	0.980	0.982	0.981	0.980	0.977	0.974	0.967
	SD	0.030	0.029	0.032	0.035	0.037	0.040	0.048
F2	Mean	0.981	0.983	0.983	0.981	0.979	0.976	0.969
	SD	0.028	0.026	0.029	0.032	0.034	0.037	0.045
LR	Mean	0.982	0.983	0.982	0.980	0.978	0.976	0.970
	SD	0.028	0.027	0.030	0.034	0.037	0.038	0.044
ST	Mean	0.986	0.984	0.968	0.984	0.984	0.981	0.965
	SD	0.019	0.017	0.032	0.024	0.025	0.028	0.037
χ^2_{FI}	Mean	0.876	0.947	0.963	0.958	0.954	0.961	0.965
	SD	0.082	0.031	0.042	0.059	0.066	0.048	0.041
χ^2_{GD}	Mean	0.989	0.990	0.989	0.988	0.986	0.984	0.979
	SD	0.018	0.017	0.018	0.021	0.022	0.025	0.030

Table A29: Mean and SD of Power Conditional on Statistic and θ for NL6 Change Pattern

		θ					
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5
F1	Mean	0.996	0.996	0.996	0.996	0.996	0.994
	SD	0.006	0.006	0.006	0.007	0.008	0.010
F2	Mean	0.996	0.997	0.997	0.997	0.996	0.994
	SD	0.006	0.005	0.006	0.006	0.007	0.009
LR	Mean	0.997	0.997	0.997	0.996	0.996	0.995
	SD	0.006	0.005	0.006	0.007	0.007	0.008
ST	Mean	0.997	0.996	0.998	0.998	0.997	0.996
	SD	0.004	0.003	0.004	0.004	0.005	0.006
χ^2_{FI}	Mean	0.960	0.975	0.980	0.980	0.989	0.990
	SD	0.012	0.013	0.017	0.025	0.021	0.015
χ^2_{GD}	Mean	0.998	0.998	0.998	0.998	0.998	0.997
	SD	0.004	0.003	0.003	0.004	0.004	0.005

Table A30: Mean and SD of Power Conditional Discrimination and Change Patterns

		High	Mediun	Low
L1	Mean	0.892	0.660	0.381
	SD	0.053	0.064	0.031
L2	Mean	0.996	0.995	0.922
	SD	0.004	0.002	0.014
L3	Mean	0.994	0.996	0.997
	SD	0.008	0.004	0.0004
NL1	Mean	0.472	0.285	0.164
	SD	0.045	0.020	0.009
NL2	Mean	0.948	0.795	0.481
	SD	0.030	0.042	0.035
NL3	Mean	0.984	0.979	0.829
	SD	0.017	0.006	0.025
NL4	Mean	0.991	0.958	0.727
	SD	0.007	0.012	0.034
NL5	Mean	0.991	0.994	0.938
	SD	0.013	0.002	0.011
NL6	Mean	0.997	0.997	0.987
	SD	0.003	0.003	0.001

Table A31: Mean and SD of Type I error Conditional Discrimination and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	Mean	0.070	0.066	0.069	0.067	0.067	0.069	0.067	0.069	0.063
	SD	0.007	0.000	0.004	0.003	0.001	0.001	0.006	0.005	0.005
Medium	Mean	0.070	0.065	0.066	0.063	0.066	0.070	0.066	0.067	0.064
	SD	0.008	0.003	0.005	0.002	0.002	0.003	0.004	0.002	0.006
Low	Mean	0.067	0.070	0.065	0.065	0.066	0.063	0.061	0.063	0.063
	SD	0.004	0.001	0.001	0.001	0.006	0.003	0.001	0.004	0.004

Table A32: Mean and SD of Power Conditional Discrimination and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	Mean	0.895	0.919	0.927	0.928	0.934	0.928	0.905	0.847	0.895
	SD	0.011	0.000	0.010	0.013	0.010	0.015	0.029	0.021	0.069
Medium	Mean	0.836	0.854	0.867	0.869	0.861	0.841	0.816	0.727	0.512
	SD	0.001	0.009	0.009	0.008	0.013	0.029	0.038	0.021	0.030
Low	Mean	0.717	0.727	0.731	0.729	0.721	0.675	0.617	0.479	0.292
	SD	0.004	0.019	0.032	0.041	0.045	0.042	0.020	0.014	0.042

Table A33: Mean and SD of Power Conditional on Discrimination and θ for L1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	Mean	0.874	0.906	0.923	0.922	0.936	0.930	0.902	0.871	0.766
	SD	0.008	0.012	0.034	0.038	0.024	0.038	0.054	0.028	0.126
Medium	Mean	0.634	0.685	0.724	0.727	0.703	0.696	0.548	0.653	0.575
	SD	0.008	0.033	0.028	0.025	0.040	0.078	0.285	0.023	0.044
Low	Mean	0.383	0.398	0.413	0.411	0.401	0.382	0.367	0.351	0.319
	SD	0.017	0.021	0.042	0.055	0.060	0.053	0.029	0.000	0.055

Table A34: Mean and SD of Power Conditional on Discrimination and θ for L2 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
High	Mean	0.987	0.995	0.997	0.994	1.000	0.999	0.999
	SD	0.005	0.000	0.000	0.001	0.000	0.001	0.001
Medium	Mean	0.992	0.995	0.996	0.997	0.995	0.995	0.992
	SD	0.001	0.001	0.000	0.002	0.004	0.005	0.002
Low	Mean	0.926	0.935	0.933	0.927	0.922	0.912	0.896
	SD	0.007	0.024	0.032	0.041	0.044	0.033	0.001

Table A35: Mean and SD of Power Conditional on Discrimination and θ for L3 Change Pattern

		θ				
		– 2.0	– 1.5	– 1.0	– 0.5	0
High	Mean	0.979	0.995	0.996	0.999	1.000
	SD	0.010	0.000	0.001	0.001	0.000
Medium	Mean	0.991	0.994	0.997	1.000	0.998
	SD	0.004	0.001	0.001	0.000	0.002
Low	Mean	0.997	0.997	0.997	0.998	0.998
	SD	0.001	0.024	0.003	0.998	0.002

Table A36: Mean and SD of Power Conditional on Discrimination and θ for NL1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	Mean	0.406	0.459	0.491	0.503	0.502	0.535	0.495	0.453	0.405
	SD	0.015	0.397	0.046	0.064	0.054	0.046	0.086	0.046	0.011
Medium	Mean	0.256	0.275	0.301	0.309	0.304	0.293	0.292	0.285	0.254
	SD	0.012	0.011	0.022	0.017	0.009	0.021	0.050	0.028	0.013
Low	Mean	0.161	0.172	0.169	0.175	0.174	0.163	0.156	0.156	0.161
	SD	0.013	0.003	0.013	0.017	0.016	0.016	0.012	0.003	0.013

Table A37: Mean and SD of Power Conditional on Discrimination and θ for NL2 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	Mean	0.933	0.954	0.966	0.959	0.975	0.969	0.959	0.943	0.876
	SD	0.015	0.003	0.015	0.018	0.007	0.021	0.036	0.028	0.070
Medium	Mean	0.754	0.801	0.838	0.837	0.817	0.811	0.808	0.783	0.706
	SD	0.002	0.029	0.025	0.016	0.030	0.070	0.078	0.034	0.031
Low	Mean	0.480	0.506	0.515	0.516	0.504	0.484	0.469	0.449	0.408
	SD	0.026	0.024	0.046	0.065	0.071	0.070	0.051	0.003	0.058

Table A38: Mean and SD of Power Conditional on Discrimination and θ for NL3 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
High	Mean	0.944	0.979	0.989	0.985	0.998	0.995	0.994	0.987
	SD	0.025	0.009	0.000	0.008	0.002	0.003	0.009	0.014
Medium	Mean	0.967	0.980	0.984	0.985	0.978	0.979	0.982	0.976
	SD	0.002	0.002	0.000	0.001	0.011	0.020	0.017	0.012
Low	Mean	0.835	0.848	0.854	0.849	0.835	0.822	0.811	0.778
	SD	0.008	0.032	0.050	0.062	0.074	0.064	0.033	0.030

Table A39: Mean and SD of Power Conditional on Discrimination and θ for NL4 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
High	Mean	0.979	0.993	0.995	0.997	0.996	0.994	0.993	0.981
	SD	0.009	0.001	0.002	0.001	0.002	0.006	0.008	0.008
Medium	Mean	0.945	0.965	0.970	0.968	0.961	0.959	0.959	0.936
	SD	0.006	0.009	0.004	0.007	0.022	0.032	0.025	0.007
Low	Mean	0.737	0.750	0.760	0.754	0.737	0.719	0.698	0.661
	SD	0.002	0.038	0.062	0.077	0.086	0.071	0.019	0.041

Table A40: Mean and SD of Power Conditional on Discrimination and θ for NL5 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
High	Mean	0.964	0.991	0.989	0.998	0.999	0.998	0.995
	SD	0.014	0.002	0.006	0.000	0.000	0.002	0.007
Medium	Mean	0.990	0.994	0.996	0.996	0.995	0.996	0.994
	SD	0.001	0.001	0.000	0.002	0.004	0.006	0.004
Low	Mean	0.943	0.949	0.947	0.942	0.935	0.933	0.919
	SD	0.010	0.021	0.027	0.040	0.038	0.025	0.000

Table A41: Mean and SD of Power Conditional on Discrimination and θ for NL6 Change Pattern

		θ					
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5
High	Mean	0.992	0.996	0.997	0.997	1.000	1.000
	SD	0.001	0.001	0.001	0.001	0.000	0.001
Medium	Mean	0.993	0.994	0.998	0.999	1.000	0.999
	SD	0.004	0.001	0.001	0.000	0.000	0.001
Low	Mean	0.987	0.988	0.988	0.986	0.986	0.984
	SD	0.005	0.006	0.010	0.013	0.011	0.006

Table A42: Mean and SD of Power Conditional on Information and Change Pattern

		Flat	Peaked
L1	Mean	0.635	0.664
	SD	0.021	0.070
L2	Mean	0.964	0.977
	SD	0.004	0.008
L3	Mean	0.996	0.996
	SD	0.003	0.006
NL1	Mean	0.294	0.320
	SD	0.014	0.036
NL2	Mean	0.728	0.755
	SD	0.019	0.056
NL3	Mean	0.921	0.940
	SD	0.007	0.019
NL4	Mean	0.879	0.904
	SD	0.011	0.027
NL5	Mean	0.969	0.980
	SD	0.003	0.009
NL6	Mean	0.992	0.995
	SD	0.001	0.003

Table A43: Mean and SD of Type I Error Conditional on Information and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Flat	Mean	0.073	0.068	0.067	0.065	0.068	0.069	0.067	0.069	0.067
	SD	0.003	0.003	0.005	0.004	0.002	0.004	0.005	0.003	0.001
Peaked	Mean	0.064	0.066	0.068	0.065	0.065	0.066	0.062	0.064	0.060
	SD	0.001	0.003	0.002	0.001	0.003	0.005	0.001	0.003	0.000

Table A44: Mean and SD of Power Conditional on Information and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Flat	Mean	0.820	0.827	0.830	0.827	0.823	0.794	0.759	0.678	0.529
	SD	0.093	0.104	0.109	0.114	0.010	0.138	0.144	0.174	0.205
Peaked	Mean	0.812	0.840	0.854	0.856	0.855	0.835	0.800	0.691	0.462
	SD	0.089	0.091	0.092	0.091	0.013	0.119	0.152	0.202	0.187

Table A45: Mean and SD of Power Conditional on Information and θ for L1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Flat	Mean	0.634	0.648	0.662	0.659	0.651	0.629	0.611	0.613	0.606
	SD	0.242	0.258	0.261	0.265	0.281	0.280	0.259	0.251	0.248
Peaked	Mean	0.626	0.679	0.711	0.715	0.709	0.709	0.692	0.637	0.501
	SD	0.249	0.252	0.254	0.251	0.255	0.271	0.281	0.271	0.202

Table A46: Mean and SD of Power Conditional on Information and θ for L2 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
Flat	Mean	0.968	0.969	0.968	0.962	0.961	0.959	0.962
	SD	0.041	0.045	0.050	0.055	0.061	0.061	0.056
Peaked	Mean	0.969	0.980	0.983	0.983	0.984	0.978	0.963
	SD	0.032	0.025	0.023	0.023	0.026	0.037	0.058

Table A47: Mean and SD of Power Conditional on Information and θ for L3 Change Pattern

		θ				
		– 2.0	– 1.5	– 1.0	– 0.5	0
Flat	Mean	0.992	0.995	0.996	0.998	0.999
	SD	0.005	0.000	0.001	0.002	0.002
Peaked	Mean	0.986	0.996	0.998	1.000	0.999
	SD	0.013	0.003	0.001	0.000	0.001

Table A48: Mean and SD of Power Conditional on Information and θ for NL1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Flat	Mean	0.284	0.300	0.301	0.306	0.308	0.311	0.279	0.280	0.278
	SD	0.124	0.149	0.150	0.147	0.151	0.178	0.145	0.134	0.128
Peaked	Mean	0.265	0.304	0.339	0.352	0.345	0.350	0.349	0.316	0.261
	SD	0.122	0.143	0.174	0.183	0.180	0.200	0.196	0.164	0.130

Table A49: Mean and SD of Power Conditional on Information and θ for NL2 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Flat	Mean	0.732	0.740	0.753	0.747	0.740	0.717	0.706	0.711	0.732
	SD	0.224	0.234	0.244	0.105	0.263	0.263	0.253	0.239	0.239
Peaked	Mean	0.712	0.767	0.793	0.794	0.790	0.793	0.784	0.739	0.712
	SD	0.233	0.221	0.222	0.210	0.217	0.233	0.249	0.265	0.235

Table A50: Mean and SD of Power Conditional on Information and θ for NL3 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
Flat	Mean	0.923	0.930	0.931	0.927	0.916	0.912	0.915	0.915
	SD	0.071	0.090	0.097	0.163	0.117	0.117	0.111	0.100
Peaked	Mean	0.908	0.942	0.954	0.953	0.957	0.953	0.943	0.913
	SD	0.071	0.062	0.056	0.052	0.061	0.074	0.094	0.135

Table A51: Mean and SD of Power Conditional on Information and θ for NL4 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
Flat	Mean	0.887	0.892	0.892	0.887	0.872	0.865	0.871	0.869
	SD	0.145	0.147	0.153	0.048	0.172	0.172	0.163	0.158
Peaked	Mean	0.887	0.913	0.924	0.926	0.924	0.916	0.895	0.849
	SD	0.129	0.119	0.105	0.103	0.110	0.128	0.160	0.189

Table A52: Mean and SD of Power Conditional on Information and θ for NL5 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
Flat	Mean	0.967	0.974	0.973	0.969	0.966	0.968	0.967
	SD	0.028	0.034	0.038	0.012	0.050	0.046	0.041
Peaked	Mean	0.964	0.983	0.982	0.988	0.987	0.983	0.972
	SD	0.021	0.016	0.015	0.016	0.021	0.029	0.046

Table A53: Mean and SD of Power Conditional on Information and θ for NL6 Change Pattern

		θ					
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5
Flat	Mean	0.990	0.993	0.992	0.991	0.992	0.992
	SD	0.006	0.007	0.010	0.265	0.012	0.011
Peaked	Mean	0.991	0.993	0.997	0.997	0.998	0.996
	SD	0.001	0.002	0.002	0.002	0.003	0.007

Table A54: Mean and SD of Power Conditional on Bank Type and Change Pattern

		HF	HP	MF	MP	LF	LP
L1	Mean	0.885	0.900	0.654	0.698	0.366	0.395
	SD	0.024	0.089	0.036	0.070	0.018	0.055
L2	Mean	0.996	0.932	0.993	0.996	0.903	0.940
	SD	0.003	0.181	0.002	0.003	0.013	0.022
L3	Mean	0.995	0.993	0.997	0.995	0.996	0.999
	SD	0.005	0.012	0.003	0.005	0.001	0.001
NL1	Mean	0.448	0.497	0.275	0.296	0.160	0.168
	SD	0.029	0.067	0.015	0.031	0.008	0.016
NL2	Mean	0.945	0.952	0.776	0.815	0.462	0.500
	SD	0.015	0.051	0.033	0.060	0.023	0.064
NL3	Mean	0.985	0.982	0.973	0.985	0.805	0.854
	SD	0.011	0.025	0.008	0.008	0.022	0.046
NL4	Mean	0.991	0.991	0.948	0.968	0.699	0.755
	SD	0.004	0.011	0.013	0.014	0.024	0.060
NL5	Mean	0.992	0.989	0.993	0.996	0.922	0.955
	SD	0.009	0.017	0.002	0.004	0.011	0.018
NL6	Mean	0.997	0.997	0.998	0.997	0.980	0.993
	SD	0.003	0.003	0.001	0.004	0.003	0.003

Table A55: Mean and SD of Type I Error Conditional on Bank Type and Statistic

		HF	HP	MF	MP	LF	LP
F1	Mean	0.059	0.052	0.056	0.052	0.054	0.047
	SD	0.003	0.004	0.004	0.004	0.002	0.004
F2	Mean	0.064	0.057	0.061	0.057	0.059	0.053
	SD	0.003	0.004	0.004	0.004	0.003	0.004
LR	Mean	0.061	0.058	0.060	0.057	0.059	0.058
	SD	0.003	0.003	0.004	0.003	0.004	0.004
ST	Mean	0.089	0.082	0.084	0.080	0.080	0.078
	SD	0.006	0.010	0.006	0.010	0.003	0.012
χ^2_{FI}	Mean	0.049	0.055	0.053	0.055	0.056	0.062
	SD	0.007	0.010	0.007	0.012	0.008	0.009
χ^2_{GD}	Mean	0.096	0.088	0.094	0.088	0.092	0.083
	SD	0.005	0.006	0.006	0.006	0.003	0.007

Table A56: Mean and SD of Power Conditional on Bank Type and Statistic

		HF	HP	MF	MP	LF	LP
F1	Mean	0.884	0.880	0.782	0.792	0.601	0.614
	SD	0.067	0.117	0.113	0.140	0.143	0.183
F2	Mean	0.889	0.885	0.790	0.799	0.612	0.625
	SD	0.064	0.111	0.110	0.135	0.141	0.180
LR	Mean	0.886	0.890	0.787	0.802	0.610	0.635
	SD	0.066	0.100	0.112	0.129	0.138	0.169
ST	Mean	0.899	0.896	0.815	0.820	0.650	0.657
	SD	0.058	0.105	0.100	0.134	0.134	0.182
χ^2_{FI}	Mean	0.833	0.861	0.745	0.789	0.588	0.652
	SD	0.082	0.080	0.117	0.094	0.123	0.135
χ^2_{GL}	Mean	0.910	0.906	0.824	0.831	0.666	0.674
	SD	0.052	0.096	0.095	0.120	0.130	0.172

Table A57: Mean and SD of Type I Error Conditional on Bank Type and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
HF	Mean	0.074	0.065	0.072	0.069	0.067	0.068	0.071	0.072	0.066
	SD	0.023	0.019	0.017	0.016	0.019	0.014	0.019	0.022	0.022
HP	Mean	0.064	0.066	0.067	0.065	0.068	0.070	0.063	0.065	0.060
	SD	0.020	0.021	0.020	0.016	0.018	0.015	0.016	0.014	0.013
MF	Mean	0.076	0.067	0.063	0.062	0.068	0.072	0.069	0.068	0.068
	SD	0.017	0.020	0.018	0.012	0.014	0.017	0.020	0.019	0.015
MP	Mean	0.064	0.063	0.070	0.065	0.064	0.069	0.063	0.066	0.060
	SD	0.020	0.020	0.018	0.014	0.017	0.017	0.016	0.013	0.014
LF	Mean	0.070	0.071	0.065	0.065	0.070	0.065	0.062	0.066	0.066
	SD	0.014	0.015	0.015	0.015	0.015	0.016	0.019	0.018	0.013
LP	Mean	0.064	0.069	0.066	0.066	0.062	0.060	0.061	0.060	0.060
	SD	0.019	0.017	0.016	0.016	0.014	0.013	0.013	0.013	0.017

Table A58: Mean and SD of Power Conditional on Bank Type and θ

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
HF	Mean	0.903	0.919	0.920	0.919	0.927	0.917	0.885	0.832	0.731
	SD	0.044	0.021	0.018	0.018	0.012	0.018	0.027	0.036	0.047
HP	Mean	0.887	0.919	0.934	0.937	0.941	0.938	0.926	0.862	0.634
	SD	0.072	0.033	0.015	0.016	0.012	0.011	0.008	0.012	0.046
MF	Mean	0.836	0.848	0.861	0.863	0.852	0.820	0.789	0.712	0.533
	SD	0.033	0.025	0.017	0.015	0.022	0.032	0.032	0.033	0.045
MP	Mean	0.835	0.860	0.873	0.874	0.871	0.862	0.843	0.741	0.491
	SD	0.039	0.026	0.019	0.019	0.020	0.014	0.013	0.022	0.042
LF	Mean	0.720	0.714	0.708	0.700	0.690	0.645	0.602	0.489	0.322
	SD	0.023	0.026	0.027	0.029	0.034	0.039	0.035	0.035	0.032
LP	Mean	0.714	0.741	0.753	0.758	0.753	0.705	0.631	0.469	0.262
	SD	0.029	0.025	0.022	0.019	0.019	0.021	0.027	0.038	0.048

Table A59: Mean and SD of Power Conditional on Bank Type and θ for L1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
HF	Mean	0.880	0.898	0.899	0.895	0.919	0.903	0.864	0.851	0.855
	SD	0.040	0.034	0.030	0.041	0.017	0.034	0.047	0.055	0.042
HP	Mean	0.868	0.915	0.947	0.949	0.953	0.957	0.941	0.890	0.677
	SD	0.063	0.049	0.026	0.027	0.023	0.014	0.010	0.018	0.060
MF	Mean	0.628	0.662	0.704	0.709	0.675	0.641	0.623	0.636	0.607
	SD	0.063	0.051	0.037	0.034	0.042	0.058	0.051	0.040	0.052
MP	Mean	0.639	0.708	0.744	0.745	0.732	0.751	0.749	0.669	0.544
	SD	0.066	0.048	0.036	0.044	0.050	0.033	0.030	0.038	0.056
LF	Mean	0.396	0.383	0.383	0.372	0.359	0.344	0.346	0.351	0.358
	SD	0.037	0.040	0.041	0.038	0.042	0.052	0.044	0.038	0.034
LP	Mean	0.371	0.413	0.442	0.450	0.444	0.419	0.387	0.351	0.280
	SD	0.042	0.044	0.042	0.038	0.039	0.039	0.039	0.044	0.061

Table A60: Mean and SD of Power Conditional on Bank Type and θ for L2 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
HF	Mean	0.990	0.995	0.997	0.994	0.999	0.998	0.998
	SD	0.022	0.011	0.008	0.016	0.002	0.004	0.003
HP	Mean	0.984	0.995	0.997	0.995	1.000	1.000	1.000
	SD	0.040	0.012	0.008	0.012	0.000	0.000	0.000
MF	Mean	0.992	0.995	0.997	0.995	0.993	0.992	0.990
	SD	0.013	0.007	0.006	0.007	0.011	0.010	0.010
MP	Mean	0.991	0.994	0.996	0.998	0.998	0.999	0.993
	SD	0.019	0.011	0.007	0.004	0.002	0.001	0.004
LF	Mean	0.921	0.918	0.910	0.898	0.891	0.889	0.897
	SD	0.019	0.021	0.021	0.032	0.038	0.028	0.019
LP	Mean	0.932	0.951	0.956	0.957	0.953	0.935	0.896
	SD	0.021	0.014	0.010	0.009	0.008	0.013	0.021

Table A61: Mean and SD of Power Conditional on Bank Type and θ for L3 Change Pattern

		θ				
		– 2.0	– 1.5	– 1.0	– 0.5	0
HF	Mean	0.987	0.996	0.996	0.999	1.000
	SD	0.032	0.011	0.010	0.004	0.000
HP	Mean	0.972	0.995	0.997	1.000	1.000
	SD	0.068	0.011	0.007	0.001	0.001
MF	Mean	0.993	0.995	0.996	0.999	1.000
	SD	0.016	0.012	0.009	0.002	0.000
MP	Mean	0.988	0.994	0.998	1.000	0.997
	SD	0.029	0.016	0.006	0.000	0.001
LF	Mean	0.996	0.995	0.995	0.996	0.997
	SD	0.007	0.008	0.007	0.007	0.005
LP	Mean	0.998	0.999	1.000	0.999	0.999
	SD	0.004	0.002	0.001	0.001	0.000

Table A62: Mean and SD of Power Conditional on Bank Type and θ for NL1 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
HF	Mean	0.417	0.462	0.458	0.458	0.464	0.502	0.434	0.421	0.413
	SD	0.073	0.050	0.061	0.049	0.059	0.043	0.059	0.061	0.063
HP	Mean	0.395	0.457	0.523	0.548	0.540	0.568	0.556	0.486	0.397
	SD	0.059	0.071	0.067	0.057	0.055	0.049	0.037	0.036	0.045
MF	Mean	0.264	0.268	0.285	0.297	0.298	0.278	0.257	0.265	0.263
	SD	0.044	0.048	0.044	0.038	0.034	0.043	0.050	0.039	0.036
MP	Mean	0.248	0.283	0.316	0.321	0.311	0.308	0.327	0.305	0.245
	SD	0.050	0.052	0.044	0.040	0.047	0.047	0.037	0.034	0.036
LF	Mean	0.170	0.170	0.160	0.164	0.163	0.152	0.148	0.154	0.158
	SD	0.026	0.027	0.029	0.027	0.026	0.031	0.034	0.027	0.023
LP	Mean	0.152	0.174	0.178	0.187	0.185	0.174	0.165	0.158	0.140
	SD	0.033	0.031	0.029	0.029	0.029	0.026	0.025	0.026	0.032

Table A63: Mean and SD of Power Conditional on Bank Type and θ for NL2 Change Pattern

		θ								
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
HF	Mean	0.943	0.952	0.956	0.946	0.970	0.955	0.933	0.923	0.926
	SD	0.044	0.041	0.033	0.050	0.018	0.046	0.055	0.057	0.042
HP	Mean	0.923	0.956	0.977	0.972	0.980	0.984	0.984	0.963	0.826
	SD	0.077	0.060	0.029	0.035	0.024	0.020	0.005	0.006	0.037
MF	Mean	0.756	0.781	0.821	0.826	0.796	0.762	0.753	0.759	0.728
	SD	0.065	0.058	0.036	0.029	0.039	0.062	0.053	0.040	0.053
MP	Mean	0.753	0.822	0.856	0.848	0.838	0.861	0.863	0.808	0.684
	SD	0.073	0.049	0.036	0.041	0.050	0.028	0.019	0.025	0.038
LF	Mean	0.498	0.489	0.482	0.470	0.454	0.435	0.433	0.451	0.450
	SD	0.041	0.042	0.046	0.042	0.046	0.058	0.055	0.044	0.039
LP	Mean	0.462	0.523	0.547	0.562	0.554	0.534	0.505	0.447	0.367
	SD	0.061	0.054	0.046	0.043	0.040	0.037	0.036	0.040	0.053

Table A64: Mean and SD of Power Conditional on Bank Type and θ for NL3 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
HF	Mean	0.962	0.985	0.989	0.991	0.996	0.992	0.987	0.977
	SD	0.091	0.033	0.018	0.015	0.008	0.019	0.019	0.040
HP	Mean	0.926	0.972	0.989	0.979	0.999	0.997	1.000	0.997
	SD	0.178	0.064	0.016	0.037	0.003	0.008	0.001	0.001
MF	Mean	0.966	0.979	0.985	0.984	0.971	0.965	0.970	0.968
	SD	0.044	0.026	0.019	0.020	0.041	0.044	0.030	0.029
MP	Mean	0.968	0.982	0.984	0.985	0.986	0.993	0.994	0.985
	SD	0.049	0.029	0.025	0.025	0.023	0.006	0.001	0.004
LF	Mean	0.840	0.825	0.818	0.806	0.782	0.777	0.788	0.800
	SD	0.029	0.035	0.035	0.036	0.053	0.052	0.037	0.029
LP	Mean	0.829	0.871	0.889	0.893	0.888	0.867	0.835	0.757
	SD	0.044	0.035	0.027	0.022	0.018	0.017	0.022	0.035

Table A65: Mean and SD of Power Conditional on Bank Type and θ for NL4 Change Pattern

		θ							
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5
HF	Mean	0.985	0.993	0.993	0.996	0.995	0.990	0.988	0.986
	SD	0.030	0.013	0.011	0.005	0.011	0.018	0.018	0.014
HP	Mean	0.973	0.992	0.996	0.997	0.998	0.999	0.999	0.975
	SD	0.061	0.017	0.006	0.006	0.005	0.003	0.000	0.014
MF	Mean	0.941	0.959	0.967	0.964	0.945	0.936	0.941	0.931
	SD	0.039	0.020	0.012	0.018	0.037	0.036	0.024	0.030
MP	Mean	0.949	0.972	0.972	0.973	0.976	0.981	0.976	0.940
	SD	0.037	0.020	0.020	0.022	0.015	0.004	0.005	0.012
LF	Mean	0.736	0.724	0.716	0.700	0.676	0.668	0.685	0.690
	SD	0.035	0.039	0.038	0.041	0.054	0.054	0.041	0.035
LP	Mean	0.738	0.777	0.804	0.808	0.798	0.769	0.711	0.631
	SD	0.045	0.038	0.032	0.06	0.026	0.027	0.034	0.045

Table A66: Mean and SD of Power Conditional on Bank Type and θ for NL5 Change Pattern

		θ						
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0
HF	Mean	0.974	0.993	0.994	0.998	0.999	0.997	0.990
	SD	0.063	0.014	0.012	0.004	0.002	0.006	0.022
HP	Mean	0.954	0.990	0.985	0.998	1.000	1.000	1.000
	SD	0.112	0.020	0.035	0.004	0.001	0.001	0.000
MF	Mean	0.991	0.995	0.996	0.995	0.992	0.992	0.992
	SD	0.020	0.010	0.007	0.010	0.018	0.016	0.013
MP	Mean	0.989	0.994	0.996	0.997	0.998	1.000	0.997
	SD	0.026	0.014	0.009	0.006	0.004	0.000	0.002
LF	Mean	0.936	0.934	0.929	0.913	0.908	0.915	0.919
	SD	0.024	0.024	0.024	0.038	0.044	0.026	0.018
LP	Mean	0.951	0.965	0.966	0.970	0.962	0.950	0.919
	SD	0.024	0.016	0.014	0.007	0.007	0.009	0.016

Table A67: Mean and SD of Power Conditional on Bank Type and θ for NL6 Change Pattern

		θ					
		– 2.0	– 1.5	– 1.0	– 0.5	0	0.5
HF	Mean	0.993	0.997	0.996	0.996	1.000	0.999
	SD	0.017	0.006	0.009	0.009	0.001	0.001
HP	Mean	0.992	0.995	0.998	0.997	1.000	1.000
	SD	0.019	0.008	0.004	0.007	0.001	0.000
MF	Mean	0.996	0.997	0.998	0.999	0.999	0.998
	SD	0.010	0.006	0.004	0.002	0.002	0.003
MP	Mean	0.990	0.993	0.998	1.000	1.000	1.000
	SD	0.024	0.018	0.006	0.001	0.000	0.000
LF	Mean	0.983	0.984	0.980	0.977	0.978	0.980
	SD	0.011	0.012	0.017	0.023	0.016	0.011
LP	Mean	0.991	0.992	0.995	0.995	0.994	0.989
	SD	0.010	0.006	0.003	0.001	0.001	0.003

Table A68: Mean and SD of Type I Error Conditional on Discrimination, Statistic and θ

			θ								
			– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	F1	Mean	0.057	0.054	0.058	0.057	0.056	0.056	0.055	0.056	0.050
		SD	0.010	0.002	0.003	0.004	0.001	0.001	0.008	0.007	0.007
	F2	Mean	0.062	0.059	0.063	0.061	0.061	0.061	0.060	0.062	0.055
		SD	0.010	0.002	0.004	0.005	0.000	0.000	0.008	0.008	0.007
	LR	Mean	0.062	0.058	0.061	0.058	0.060	0.062	0.058	0.060	0.055
		SD	0.003	0.000	0.003	0.002	0.000	0.000	0.004	0.004	0.002
	ST	Mean	0.093	0.089	0.090	0.084	0.086	0.081	0.083	0.084	0.076
		SD	0.009	0.006	0.001	0.002	0.000	0.003	0.007	0.013	0.021
	χ^2_{FI}	Mean	0.052	0.044	0.050	0.051	0.047	0.062	0.051	0.054	0.058
		SD	0.008	0.002	0.009	0.001	0.005	0.000	0.001	0.012	0.023
	χ^2_{GD}	Mean	0.092	0.090	0.094	0.091	0.094	0.093	0.095	0.096	0.086
		SD	0.015	0.001	0.005	0.004	0.003	0.003	0.008	0.010	0.014
Medium	F1	Mean	0.057	0.053	0.054	0.052	0.054	0.059	0.054	0.053	0.050
		SD	0.008	0.003	0.004	0.002	0.003	0.002	0.007	0.003	0.008
	F2	Mean	0.062	0.058	0.059	0.057	0.058	0.065	0.059	0.059	0.055
		SD	0.009	0.003	0.004	0.002	0.002	0.003	0.007	0.004	0.008
	LR	Mean	0.062	0.056	0.059	0.056	0.058	0.062	0.057	0.058	0.056
		SD	0.007	0.002	0.005	0.002	0.002	0.003	0.003	0.001	0.004
	ST	Mean	0.092	0.088	0.087	0.076	0.080	0.086	0.083	0.079	0.071
		SD	0.002	0.000	0.007	0.004	0.002	0.001	0.008	0.011	0.016
	χ^2_{FI}	Mean	0.052	0.044	0.049	0.054	0.055	0.055	0.050	0.060	0.067
		SD	0.012	0.004	0.006	0.000	0.009	0.004	0.007	0.016	0.014
	χ^2_{GD}	Mean	0.094	0.092	0.092	0.084	0.091	0.096	0.093	0.091	0.084
		SD	0.010	0.005	0.004	0.004	0.000	0.003	0.007	0.007	0.012

Table A68 – Continued on the next page.

Table A68 (continued): Mean and SD of Type I Error Conditional on Discrimination, Statistic and θ

			θ								
			– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Low	F1	Mean	0.050	0.053	0.051	0.051	0.052	0.051	0.049	0.048	0.047
		SD	0.010	0.004	0.002	0.012	0.010	0.005	0.004	0.006	0.009
	F2	Mean	0.055	0.058	0.057	0.057	0.059	0.056	0.054	0.054	0.053
		SD	0.011	0.004	0.001	0.014	0.012	0.005	0.003	0.006	0.009
	LR	Mean	0.063	0.064	0.058	0.058	0.058	0.055	0.053	0.056	0.057
		SD	0.001	0.001	0.003	0.017	0.015	0.002	0.001	0.002	0.002
	ST	Mean	0.088	0.087	0.081	0.080	0.078	0.072	0.074	0.074	0.067
		SD	0.010	0.007	0.006	0.019	0.017	0.007	0.010	0.015	0.013
	χ^2_{FI}	Mean	0.062	0.065	0.057	0.057	0.059	0.054	0.053	0.062	0.074
		SD	0.002	0.001	0.001	0.022	0.023	0.006	0.017	0.014	0.016
	χ^2_{GD}	Mean	0.083	0.091	0.088	0.090	0.089	0.087	0.086	0.086	0.081
		SD	0.015	0.006	0.001	0.016	0.018	0.006	0.007	0.011	0.009

Table A69: Mean and SD of Power Conditional on Discrimination, Statistic and θ

			θ								
			– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
High	F1	Mean	0.909	0.922	0.928	0.930	0.931	0.925	0.904	0.841	0.651
		SD	0.005	0.003	0.011	0.013	0.011	0.014	0.021	0.006	0.101
	F2	Mean	0.912	0.925	0.931	0.933	0.934	0.929	0.907	0.848	0.665
		SD	0.005	0.003	0.010	0.013	0.011	0.013	0.021	0.006	0.095
	LR	Mean	0.913	0.925	0.930	0.931	0.933	0.928	0.905	0.848	0.680
		SD	0.001	0.004	0.012	0.013	0.011	0.013	0.024	0.015	0.067
	ST	Mean	0.928	0.936	0.933	0.934	0.945	0.939	0.913	0.854	0.697
		SD	0.002	0.004	0.006	0.009	0.009	0.012	0.028	0.024	0.104
	χ^2_{FI}	Mean	0.779	0.866	0.895	0.896	0.914	0.902	0.878	0.816	0.680
		SD	0.052	0.018	0.014	0.016	0.009	0.026	0.062	0.074	0.042
	χ^2_{GD}	Mean	0.930	0.940	0.945	0.946	0.947	0.943	0.925	0.875	0.721
		SD	0.004	0.002	0.009	0.012	0.009	0.010	0.018	0.004	0.090
Medium	F1	Mean	0.834	0.849	0.860	0.862	0.857	0.835	0.805	0.705	0.477
		SD	0.000	0.009	0.010	0.010	0.013	0.024	0.031	0.013	0.049
	F2	Mean	0.840	0.855	0.865	0.867	0.862	0.841	0.812	0.716	0.493
		SD	0.000	0.008	0.010	0.010	0.012	0.023	0.031	0.012	0.047
	LR	Mean	0.840	0.855	0.866	0.866	0.860	0.837	0.810	0.718	0.500
		SD	0.001	0.010	0.010	0.009	0.012	0.027	0.036	0.019	0.029
	ST	Mean	0.865	0.877	0.884	0.884	0.879	0.861	0.835	0.743	0.530
		SD	0.003	0.009	0.009	0.009	0.014	0.022	0.026	0.002	0.066
	χ^2_{FI}	Mean	0.768	0.808	0.839	0.842	0.825	0.804	0.790	0.714	0.512
		SD	0.010	0.007	0.005	0.000	0.016	0.060	0.077	0.071	0.060
	χ^2_{GD}	Mean	0.867	0.880	0.889	0.890	0.886	0.868	0.844	0.763	0.561
		SD	0.000	0.009	0.008	0.008	0.011	0.020	0.026	0.010	0.047

Table A69 – Continued on the next page.

Table A69 (Continued): Mean and SD of Power Conditional on Discrimination, Statistic and θ

			θ								
			– 2.0	– 1.5	– 1.0	– 0.5	0	0.5	1.0	1.5	2.0
Low	F1	Mean	0.694	0.709	0.713	0.713	0.706	0.655	0.588	0.440	0.247
		SD	0.011	0.015	0.030	0.038	0.039	0.034	0.014	0.021	0.056
	F2	Mean	0.704	0.718	0.723	0.721	0.715	0.665	0.600	0.455	0.262
		SD	0.009	0.015	0.029	0.038	0.038	0.035	0.014	0.020	0.055
	LR	Mean	0.714	0.721	0.723	0.721	0.712	0.663	0.603	0.465	0.283
		SD	0.002	0.021	0.033	0.041	0.043	0.042	0.021	0.010	0.031
	ST	Mean	0.747	0.754	0.756	0.751	0.744	0.699	0.637	0.492	0.301
		SD	0.008	0.024	0.030	0.037	0.035	0.024	0.003	0.043	0.064
	χ^2_{FI}	Mean	0.691	0.698	0.703	0.703	0.694	0.650	0.611	0.499	0.332
		SD	0.008	0.023	0.042	0.062	0.080	0.088	0.066	0.034	0.015
	χ^2_{GD}	Mean	0.749	0.762	0.765	0.764	0.758	0.716	0.659	0.523	0.329
		SD	0.008	0.015	0.026	0.033	0.034	0.030	0.010	0.027	0.062

Table A70 through-Table A88: Proportion Agreement Under Various Conditions

Table A70: Proportion Agreement Between Methods for Type I Error and Power

	Type I Error						Power					
	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.99	0.97	0.98	0.96	1.00	0.99	0.99	0.97	0.95	0.97
F2		1.00	0.99	0.97	0.98	0.97		1.00	0.99	0.98	0.95	0.97
LR			1.00	0.98	0.98	0.97			1.00	0.98	0.95	0.97
ST				1.00	0.97	0.99				1.00	0.94	0.99
χ^2_{FI}					1.00	0.96					1.00	0.94
χ^2_{GD}						1.00						1.00

Table A71: Proportion Agreement Between Methods for L1 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.97	0.94	0.92	0.92
F2		1.00	0.98	0.95	0.92	0.94
LR			1.00	0.96	0.94	0.94
ST				1.00	0.91	0.98
χ^2_{FI}					1.00	0.91
χ^2_{GD}						1.00

Table A72: Proportion Agreement Between Methods for L2 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	1.00	1.00	0.99	0.97	0.99
F2		1.00	1.00	0.99	0.97	0.99
LR			1.00	0.99	0.97	0.99
ST				1.00	0.97	1.00
χ^2_{FI}					1.00	0.97
χ^2_{GD}						1.00

Table A73: Proportion Agreement Between Methods for L3 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	1.00	0.99	1.00	0.98	1.00
F2		1.00	1.00	1.00	0.98	1.00
LR			1.00	1.00	0.98	1.00
ST				1.00	0.98	1.00
χ^2_{FI}					1.00	0.98
χ^2_{GD}						1.00

Table A74: Proportion Agreement Between Methods for NL1 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.97	0.93	0.93	0.91
F2		1.00	0.98	0.94	0.94	0.93
LR			1.00	0.94	0.95	0.93
ST				1.00	0.91	0.97
χ^2_{FI}					1.00	0.91
χ^2_{GD}						1.00

Table A75: Proportion Agreement Between Methods for NL2 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.98	0.95	0.91	0.94
F2		1.00	0.98	0.96	0.91	0.95
LR			1.00	0.96	0.92	0.95
ST				1.00	0.90	0.98
χ^2_{FI}					1.00	0.90
χ^2_{GD}						1.00

Table A76: Proportion Agreement Between Methods for NL3 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	1.00	0.99	0.97	0.93	0.98
F2		1.00	0.99	0.98	0.93	0.98
LR			1.00	0.98	0.94	0.98
ST				1.00	0.92	0.99
χ^2_{FI}					1.00	0.93
χ^2_{GD}						1.00

Table A77: Proportion Agreement Between Methods for NL4 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	0.99	0.99	0.97	0.94	0.97
F2		1.00	0.99	0.98	0.95	0.97
LR			1.00	0.98	0.95	0.97
ST				1.00	0.94	0.99
χ^2_{FI}					1.00	0.94
χ^2_{GD}						1.00

Table A78: Proportion Agreement Between Methods for NL5 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	1.00	1.00	0.99	0.96	0.99
F2		1.00	1.00	0.99	0.96	0.99
LR			1.00	0.99	0.96	0.99
ST				1.00	0.96	0.99
χ^2_{FI}					1.00	0.96
χ^2_{GD}						1.00

Table A79: Proportion Agreement Between Methods for NL6 Change Pattern

	F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
F1	1.00	1.00	1.00	1.00	0.98	1.00
F2		1.00	1.00	1.00	0.98	1.00
LR			1.00	1.00	0.98	1.00
ST				1.00	0.98	1.00
χ^2_{FI}					1.00	0.98
χ^2_{GD}						1.00

Table A80: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = -2$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.90	0.98	HP	1.00	1.00	0.99	0.97	0.83	0.98
	F2		1.00	0.99	0.98	0.90	0.98			1.00	0.99	0.97	0.83	0.98
	LR			1.00	0.98	0.91	0.98				1.00	0.98	0.84	0.98
	ST				1.00	0.89	0.99					1.00	0.83	0.99
	χ^2_{FI}					1.00	0.89						1.00	0.83
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.97	0.93	0.97	MP	1.00	0.99	0.99	0.96	0.91	0.97
	F2		1.00	0.99	0.97	0.93	0.97			1.00	0.99	0.97	0.91	0.97
	LR			1.00	0.97	0.93	0.97				1.00	0.97	0.92	0.97
	ST				1.00	0.92	0.99					1.00	0.90	0.99
	χ^2_{FI}					1.00	0.91						1.00	0.90
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.98	0.96	0.95	0.95	LP	1.00	0.99	0.97	0.93	0.93	0.95
	F2		1.00	0.99	0.97	0.95	0.96			1.00	0.98	0.94	0.93	0.96
	LR			1.00	0.97	0.96	0.96				1.00	0.96	0.95	0.97
	ST				1.00	0.95	0.98					1.00	0.93	0.98
	χ^2_{FI}					1.00	0.94						1.00	0.94
	χ^2_{GD}						1.00							1.00

Table A81: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = -1.5$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.95	0.98	HP	1.00	1.00	0.99	0.97	0.92	0.98
	F2		1.00	0.99	0.98	0.95	0.98			1.00	0.99	0.98	0.92	0.98
	LR			1.00	0.98	0.95	0.98				1.00	0.98	0.93	0.98
	ST				1.00	0.94	0.99					1.00	0.91	0.99
	χ^2_{FI}					1.00	0.94						1.00	0.91
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.97	0.95	0.97	MP	1.00	1.00	0.99	0.97	0.94	0.97
	F2		1.00	0.99	0.97	0.95	0.97			1.00	0.99	0.97	0.94	0.97
	LR			1.00	0.97	0.95	0.97				1.00	0.98	0.95	0.98
	ST				1.00	0.93	0.99					1.00	0.93	0.99
	χ^2_{FI}					1.00	0.93						1.00	0.93
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.99	0.96	0.95	0.95	LP	1.00	0.99	0.98	0.95	0.95	0.95
	F2		1.00	0.99	0.97	0.95	0.96			1.00	0.99	0.96	0.95	0.96
	LR			1.00	0.97	0.96	0.96				1.00	0.97	0.96	0.96
	ST				1.00	0.94	0.99					1.00	0.94	0.99
	χ^2_{FI}					1.00	0.93						1.00	0.94
	χ^2_{GD}						1.00							1.00

Table A82: Proportion Agreement Between Methods Conditional on Bank Type at $\boldsymbol{\theta = -1.0}$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.96	0.98	HP	1.00	1.00	0.99	0.97	0.96	0.98
	F2		1.00	0.99	0.98	0.96	0.98			1.00	0.99	0.97	0.96	0.99
	LR			1.00	0.98	0.96	0.98				1.00	0.97	0.97	0.98
	ST				1.00	0.95	0.99					1.00	0.94	0.98
	χ^2_{FI}					1.00	0.95						1.00	0.95
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.97	0.96	0.97	MP	1.00	1.00	0.99	0.97	0.96	0.97
	F2		1.00	0.99	0.98	0.97	0.97			1.00	0.99	0.98	0.96	0.98
	LR			1.00	0.98	0.97	0.97				1.00	0.98	0.96	0.98
	ST				1.00	0.96	0.99					1.00	0.95	0.99
	χ^2_{FI}					1.00	0.95						1.00	0.95
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.99	0.96	0.95	0.95	LP	1.00	0.99	0.98	0.96	0.96	0.95
	F2		1.00	0.99	0.97	0.96	0.96			1.00	0.99	0.97	0.96	0.96
	LR			1.00	0.97	0.96	0.95				1.00	0.97	0.97	0.96
	ST				1.00	0.94	0.99					1.00	0.95	0.99
	χ^2_{FI}					1.00	0.93						1.00	0.95
	χ^2_{GD}						1.00							1.00

Table A83: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = -0.5$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.96	0.98	HP	1.00	1.00	0.99	0.97	0.96	0.98
	F2		1.00	0.99	0.98	0.96	0.98			1.00	0.99	0.98	0.96	0.99
	LR			1.00	0.98	0.96	0.98				1.00	0.98	0.96	0.98
	ST				1.00	0.95	0.99					1.00	0.95	0.99
	χ^2_{FI}					1.00	0.95						1.00	0.95
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.98	0.96	0.97	MP	1.00	1.00	0.99	0.98	0.96	0.97
	F2		1.00	0.99	0.98	0.97	0.98			1.00	0.99	0.98	0.96	0.98
	LR			1.00	0.98	0.97	0.97				1.00	0.98	0.96	0.98
	ST				1.00	0.96	0.99					1.00	0.95	0.99
	χ^2_{FI}					1.00	0.96						1.00	0.95
	χ^2_{GD}						1.00							0.99
LF	F1	1.00	0.99	0.99	0.96	0.94	0.95	LP	1.00	0.99	0.99	0.96	0.97	0.95
	F2		1.00	0.99	0.97	0.95	0.95			1.00	0.99	0.97	0.97	0.96
	LR			1.00	0.97	0.95	0.95				1.00	0.97	0.98	0.96
	ST				1.00	0.94	0.98					1.00	0.97	0.99
	χ^2_{FI}					1.00	0.92						1.00	0.96
	χ^2_{GD}					0.92	1.00							1.00

Table A84: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = 0$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.97	0.98	HP	1.00	1.00	1.00	0.99	0.97	0.98
	F2		1.00	1.00	0.99	0.97	0.98			1.00	1.00	0.99	0.97	0.99
	LR			1.00	0.99	0.98	0.98				1.00	0.99	0.98	0.99
	ST				1.00	0.97	1.00					1.00	0.97	1.00
	χ^2_{FI}					1.00	0.96						1.00	0.97
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.98	0.95	0.97	MP	1.00	1.00	0.99	0.98	0.96	0.97
	F2		1.00	0.99	0.98	0.95	0.98			1.00	0.99	0.98	0.96	0.98
	LR			1.00	0.98	0.95	0.97				1.00	0.98	0.96	0.97
	ST				1.00	0.94	0.99					1.00	0.95	0.99
	χ^2_{FI}					1.00	0.94						1.00	0.94
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.98	0.96	0.94	0.95	LP	1.00	0.99	0.99	0.97	0.98	0.95
	F2		1.00	0.99	0.97	0.94	0.95			1.00	0.99	0.97	0.98	0.96
	LR			1.00	0.96	0.94	0.95				1.00	0.97	0.98	0.96
	ST				1.00	0.92	0.98					1.00	0.98	0.99
	χ^2_{FI}					1.00	0.91						1.00	0.97
	χ^2_{GD}						1.00							1.00

Table A85: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = 0.5$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.98	0.95	0.98	HP	1.00	1.00	1.00	0.99	0.98	0.98
	F2		1.00	0.99	0.99	0.95	0.98			1.00	1.00	0.99	0.98	0.98
	LR			1.00	0.99	0.96	0.98				1.00	0.99	0.98	0.98
	ST				1.00	0.95	0.99					1.00	0.97	1.00
	χ^2_{FI}					1.00	0.95						1.00	0.97
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.97	0.94	0.96	MP	1.00	0.99	0.99	0.97	0.98	0.97
	F2		1.00	0.99	0.98	0.93	0.97			1.00	0.99	0.98	0.98	0.97
	LR			1.00	0.97	0.94	0.96				1.00	0.98	0.98	0.97
	ST				1.00	0.92	0.99					1.00	0.97	0.99
	χ^2_{FI}					1.00	0.91						1.00	0.96
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.98	0.95	0.95	0.94	LP	1.00	0.99	0.99	0.97	0.97	0.94
	F2		1.00	0.98	0.96	0.94	0.95			1.00	0.99	0.97	0.98	0.95
	LR			1.00	0.96	0.95	0.94				1.00	0.98	0.98	0.96
	ST				1.00	0.91	0.99					1.00	0.98	0.98
	χ^2_{FI}					1.00	0.90						1.00	0.97
	χ^2_{GD}						1.00							1.00

Table A86: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = 1.0$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	1.00	0.99	0.97	0.94	0.97	HP	1.00	1.00	0.99	0.98	0.99	0.98
	F2		1.00	0.99	0.97	0.94	0.98			1.00	1.00	0.99	0.99	0.98
	LR			1.00	0.97	0.95	0.97					0.99	0.99	0.98
	ST				1.00	0.92	0.98					1.00	0.99	0.99
	χ^2_{FI}					1.00	0.92						1.00	0.98
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.99	0.96	0.95	0.96	MP	1.00	0.99	0.99	0.97	0.98	0.96
	F2		1.00	0.99	0.97	0.94	0.96			1.00	0.99	0.98	0.98	0.97
	LR			1.00	0.97	0.95	0.96				1.00	0.98	0.99	0.97
	ST				1.00	0.92	0.99					1.00	0.99	0.99
	χ^2_{FI}					1.00	0.91						1.00	0.98
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.98	0.94	0.96	0.93	LP	1.00	0.99	0.98	0.96	0.95	0.94
	F2		1.00	0.99	0.95	0.95	0.94			1.00	0.99	0.97	0.96	0.95
	LR			1.00	0.95	0.96	0.94				1.00	0.98	0.96	0.95
	ST				1.00	0.93	0.99					1.00	0.97	0.97
	χ^2_{FI}					1.00	0.92						1.00	0.98
	χ^2_{GD}						1.00							1.00

Table A87: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = 1.5$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	0.99	0.98	0.95	0.91	0.96	HP	1.00	0.99	0.98	0.97	0.97	0.96
	F2		1.00	0.98	0.96	0.91	0.96			1.00	0.99	0.98	0.98	0.96
	LR			1.00	0.95	0.92	0.96				1.00	0.98	0.98	0.96
	ST				1.00	0.88	0.98					1.00	0.98	0.98
	χ^2_{FI}					1.00	0.88						1.00	0.98
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.98	0.95	0.94	0.93	MP	1.00	0.99	0.98	0.97	0.95	0.94
	F2		1.00	0.98	0.96	0.94	0.95			1.00	0.98	0.97	0.96	0.95
	LR			1.00	0.96	0.95	0.94				1.00	0.98	0.97	0.96
	ST				1.00	0.92	0.98					1.00	0.98	0.97
	χ^2_{FI}					1.00	0.91						1.00	0.99
	χ^2_{GD}						1.00						0	1.00
LF	F1	1.00	0.99	0.98	0.94	0.95	0.92	LP	1.00	0.99	0.97	0.96	0.91	0.93
	F2		1.00	0.98	0.95	0.96	0.93			1.00	0.98	0.97	0.93	0.94
	LR			1.00	0.95	0.97	0.93				1.00	0.99	0.94	0.96
	ST				1.00	0.95	0.98					1.00	0.95	0.96
	χ^2_{FI}					1.00	0.94						1.00	0.98
	χ^2_{GD}						1.00							1.00

Table A88: Proportion Agreement Between Methods Conditional on Bank Type at $\theta = 2.0$

		Type I Error							Power					
		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}		F1	F2	LR	ST	χ^2_{FI}	χ^2_{GD}
HF	F1	1.00	0.99	0.98	0.94	0.92	0.94	HP	1.00	0.99	0.96	0.95	0.89	0.93
	F2		1.00	0.98	0.94	0.92	0.95			1.00	0.97	0.95	0.91	0.95
	LR			1.00	0.95	0.93	0.95				1.00	0.98	0.94	0.97
	ST				1.00	0.89	0.98					1.00	0.93	0.96
	χ^2_{FI}					1.00	0.88						1.00	0.96
	χ^2_{GD}						1.00							1.00
MF	F1	1.00	0.99	0.98	0.94	0.92	0.93	MP	1.00	0.99	0.97	0.96	0.91	0.93
	F2		1.00	0.98	0.95	0.92	0.94			1.00	0.97	0.96	0.92	0.94
	LR			1.00	0.95	0.93	0.94				1.00	0.99	0.94	0.96
	ST				1.00	0.91	0.98					1.00	0.94	0.96
	χ^2_{FI}					1.00	0.90						1.00	0.98
	χ^2_{GD}						1.00							1.00
LF	F1	1.00	0.99	0.98	0.95	0.95	0.93	LP	1.00	0.99	0.96	0.95	0.89	0.93
	F2		1.00	0.98	0.96	0.96	0.94			1.00	0.96	0.95	0.90	0.95
	LR			1.00	0.96	0.97	0.94				1.00	0.98	0.93	0.97
	ST				1.00	0.97	0.98					1.00	0.93	0.97
	χ^2_{FI}					1.00	0.95						1.00	0.95
	χ^2_{GD}						1.00							1.00